Completing Partial Reaction Equations with Rule and Language Model-based Methods

van Wijngaarden, Matthijs; Vogel, Gabriel; Weber, Jana Marie

DOI
10.1016/B978-0-443-28824-1.50524-X

Publication date
2024

Document Version
Final published version

Published in
Computer Aided Chemical Engineering

Flavio Manenti, Gintaras V. Reklaitis (Eds.), Proceedings of the 34th European Symposium on Computer Aided Process Engineering / 15th International Symposium on Process Systems Engineering (ESCAPE34/PSE24), June 2-6, 2024, Florence, Italy
 http://dx.doi.org/10.1016/B978-0-443-28824-1.50524-X

Completing Partial Reaction Equations with Rule and Language Model-based Methods

Matthijs van Wijngaarden[a], Gabriel Vogel[a], Jana Marie Weber[a,*]

[a]*Delft Bioinformatics Lab, Department of Intelligent Systems, Delft University of Technology, Van der Maasweg 9, Delft 2629 HZ, The Netherlands*
j.m.weber@tudelft.nl

Abstract

Large chemical reaction data sets often suffer from incompleteness, such as missing molecules or stoichiometric information. Incomplete chemical reaction equations currently hinder us to perform automated mass balances across large sets of chemical reactions. In this work, we integrate two approaches for computational completion of partial reaction equations. Specifically, we combine a rule-based method and a machine learning model, a tailored version of the pre-trained Molecular Transformer, to complete reactions. The rule-based method takes sets of helper species into a linear solver and therewith balances some incomplete reactions. The machine learning model is trained to take partial reactions as inputs and predicts missing molecules and stoichiometries. We apply our methodology to the USPTO STEREO chemical reaction data set. The rule-based method completes about 50 % of the reactions. The language model shows a top 1 accuracy of 88.3 % on our test set and high validity (> 99 % of outputs are valid SMILES).

Keywords: Molecular transformer, Chemical reaction completion, rule-based methods, reaction SMILES, Language models

1. Introduction

The digitalisation of patents and publications in chemical science has led to a substantial body of electronically accessible chemical reactions (Lowe, 2012). This body of reactions is a valuable data source in predictive chemistry, e.g. for reaction prediction, retrosynthesis, reaction yield prediction, or reaction condition prediction (Schwaller et al., 2019; Liu et al., 2017; Schwaller et al., 2021; Gao et al., 2018). Recently, there has also been a growing interest in using this data for automated reaction pathways selection and early-stage sustainability analysis (Ulonska et al., 2016; Weber et al., 2022). Yet, most chemical reaction databases are incomplete. For instance, one can find recorded reactions without temperature and pressure, without yield, or without information about solvents or catalysts (Jacob et al., 2017a). Also, fundamental information such as reaction equations are incomplete. They often lack co-reactants, by-products, and the stoichiometric coefficients. The lack of this knowledge currently limits automated mass balances across the large body of reaction alternatives. This is particularly relevant for mass-based assessment strategies, e.g. sustainability focused assessment, of chemical reactions (Jacob et al., 2017b; Weber et al., 2021).

To address this problem, recent works aim to curate incomplete chemical reaction equations. Vaucher et al. (2020) proposed to complete chemical reactions through a transformer-based language model (LM). Note that their definition of completeness does not correspond to mass balance complete reactions; it corresponds to predicting the original atom-wise incomplete database entry. Arun et al. (2023) developed an algorithm that balances chemical reactions by adding small “helper” molecules. This procedure is

one of eight data processing steps for impurity prediction where the balancing step is used for filtering purposes. Zhang et al. (2023) used a similar algorithm, but also included a transformer-based encoder-only LM, based on RoBERTa (Yinhan et al., 2019), that predicts missing molecules. Their hybrid approach is a prominent step towards solving the reaction completion problem and works in an iterative fashion between the rule-based and the language model-based part. We also propose a hybrid approach. Our approach works sequentially, first through a rule-based approach like the work of Zhang et al. (2023), and then through an autoregressive transformer-based encoder-decoder LM, based on the original transformer architecture (Vaswani et al., 2017).

2. Methods

Two methods are combined for completing incomplete chemical reactions in this work. We define an incomplete reaction equation as an equation in which the number of atoms and charges on the left-hand side (LHS) and right-hand-side (RHS) of an equation are not balanced with one another.

2.1. Dataset

The dataset used for this work is the publicly available patent-mined dataset known as the USPTO STEREO (https://ibm.ent.box.com/v/ReactionSeq2SeqDataset) of which 3.5 % are balanced reactions and 96.5 % are imbalanced.

2.2. Rule-based reaction completion

The rule-based method uses a set of hard-coded mathematical and chemical rules to identify missing molecules, i.e. small helper species, necessary for a balanced reaction. Additionally, stoichiometric ratios are determined. The rule-based reaction completion is solved through a linear solver.

2.2.1 Helper species selection

Different sets of helper species considered in this work are depicted in Figure 1. Set A is the strict uncharged set, set B and C make up the strict charged helper species, and set D is taken from literature (Arun et al., 2023) illustrating a more lenient selection of helper species. Here, we test the usage of single helper species first and only if the algorithm is unsuccessful, combinations of two helper species (sets A+A, A+B, A+C, A+D).

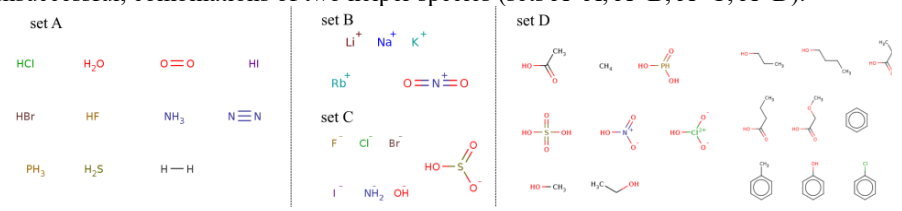

Figure 1. Helper species sets. Set A is the strict uncharged set, set B and C make up the strict charged based helper species, and set D (lenient set) is based on Arun et al. (2023).

2.2.2 Rule-based algorithm with linear solver

The rule-based algorithm can be subdivided into four parts. Firstly, atom and charge-level balances are calculated for a reaction equation, identifying the surplus or lack of atoms/charges from the left-hand-side (LHS) to the right-hand-side (RHS). Secondly, helper species are selected when their atom types coincide with the in step one identified imbalanced atom types. In the first iteration, only one helper species (single-type) is selected to complete the equation and in the second iteration of this step, a combination of two helper species (pairwise-type) is selected. Thirdly, the linear solver identifies the stoichiometry of the added helper species. For single-typed solutions, the linear solver

checks if the number of missing atoms can be divided by the number of atoms of the helper molecule. For pairwise solutions, the less ambiguous helper species is selected first: in the case of a charge imbalance, a charged helper species; without a charge imbalance, a helper species with unique atom type. The stoichiometric value of the selected molecule is then set to balance the charge or unique atom type. The atom/charge balance is updated, and the secondary helper species is selected as in the single-type solution. Lastly, if a reaction cannot be completed through the previous steps, we check if the atom imbalance exactly coincides with one of the reactants or products. If this is the case, we assume that that molecule was incorrectly added to the reactant or product side, while it should have been recorded as a reagent and thus remove it.

2.3. Language-model based reaction completion

The second method is a transformer-based encoder-decoder LM that is trained on pairs of partial and complete reactions and predicts missing molecules and stoichiometries. Molecules are presented as string-formatted words with their atoms as tokens using SMILES (simplified molecular input line entry system) notation. Reactions are a sequence of words: a sentence, see Figure 2 (a). We fine-tune the Molecular Transformer (Schwaller et al., 2019) on a reaction completion task. The averaged 20 checkpoint Molecular Transformer (https://ibm.ent.box.com/v/MolecularTransformerModels) from the USPTO STEREO dataset with separated solvents is used for initialisation. To generate fine-tuning data, we partialised the data set of complete reactions obtained from the rule-based model. We then subsequently train the model to predict the complete reaction equation from a partialised equation, see Figure 2 (b). Each reaction was first assigned to the train, test, or validate data set with a data split of 90/5/5 and then partialised. Reactions from the test set were partialised only once, while reactions belonging to training and validation set are partialised up to ten times depending on the number of possible combinations, keeping at least 50 % of the atoms from the complete reaction equation. In some cases, two different reactions produce the same partialised reaction. Then, both correct answers are recorded for each partial reaction. During testing, the prediction of either one is considered correct.

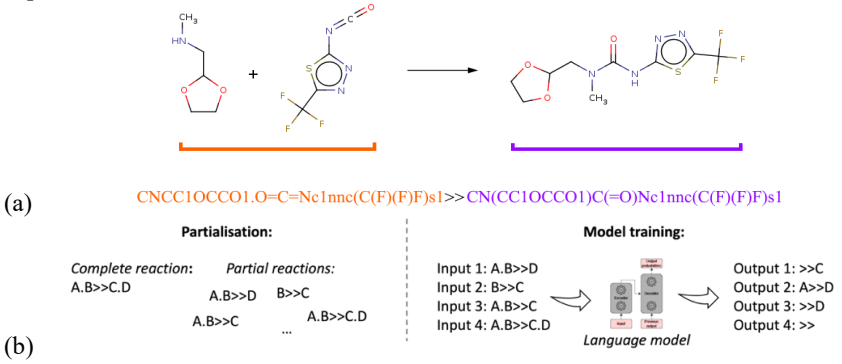

Figure 2. Illustration of a reaction SMILES as input for the LM (a) and the partialisation strategy (b). In (a), tokens before ">>" correspond to reactant molecules and tokens afterwards to product molecules. In (b), partial reactions are used for model training.

2.3 Model evaluation

We propose three evaluation scenarios. Solutions from the rule-based model are considered complete if they fulfil the atom and charge balance as also proposed by Zhang et al., (2023). Note that this is an approximation with false positives, and that for example

chemical template matching and expert judgments would be beneficial. With the previous assumption, we can consider the data completed through the rule-based approach as ground truth for the LM and can thus record the prediction accuracy. Lastly, we test the LM on the reactions that could not be completed by the rule-based approach. There, we evaluate if the model predicts atom- and charge-balanced reactions and perform an additional consistency check through the round-trip accuracy, inspired by retrosynthesis prediction tasks (Schwaller et al., 2019). We define a round-trip accurate prediction as one whose output, if newly partialised (50 - 80 % of atoms remain) and fed into the LM, leads to the same complete chemical reaction.

3. Results and discussion

3.1. Rule-based completion

Using the rule-based reaction completion algorithm increases the fraction of complete reactions from 3.5 to 49.37 % (strict helper species set) and 55.57 % (lenient helper species set). We outline the completion rate per algorithm stage for the strict species set in Table 1. Notably, the helper species-based completion algorithm contributes to most of the completed data while the erroneous reactant step only identifies very few mislabelled reaction records. Our cumulative results are in line with our reimplementation of the rule-based method ChemBalancer (Zhang et al., 2023) that resulted in a curation rate of 54 % on a sample set of incomplete reactions.

Table 1. Data completion rate at each step in the rule-based algorithm using the strict set of helper species and the cumulative value also for the strict species set.

	initial data	strict	err. reactants	cumulative
Completion rate [%]	3.5	45.53	0.37	49.37

3.2. Language-based completion with ground truth assumption

The fine-tuned Molecular Transformer shows a top five accuracy of 95.6 % on the test set of the by the rule-based method completed dataset. 99.78 % of top 1 predictions are valid SMILES outputs, which is in line with our expectations as the model was initialised on the previously trained Molecular Transformer. Yet, the validity slightly decreases in respective next n predictions. Table 2 illustrates the performance and validity.

Table 2. Performance of model on test set. BS stands for beam search and top n considers the accuracy of the first n predictions. For SMILES validity top n corresponds to the n^{th} prediction.

	top 1 BS 1	top 1 BS 5	top 2 BS 5	top 3 BS 5	top 4 BS 1	top 5 BS 1
Top-n accuracy [%]	88.3	88.8	93.6	94.8	95.3	95.6
Valid SMILES output [%]	99.78	99.88	88.07	94.23	93.27	91.94

3.2.1. Degree of partialisation

When limiting the scope of the reaction incompleteness problem to partial reactions with exactly one molecule missing, our LM achieved a top 1 accuracy of 96.3 %. In Figure 3 (a), we outline the model's performance for remaining combinations of reaction partialisation scenarios. Note that for zero missing molecules, the model always predicts that no additional molecule is needed, thus that the reaction is already complete. Furthermore, we observe a gradient from left to right highlighting the increasing complexity with more missing molecules.

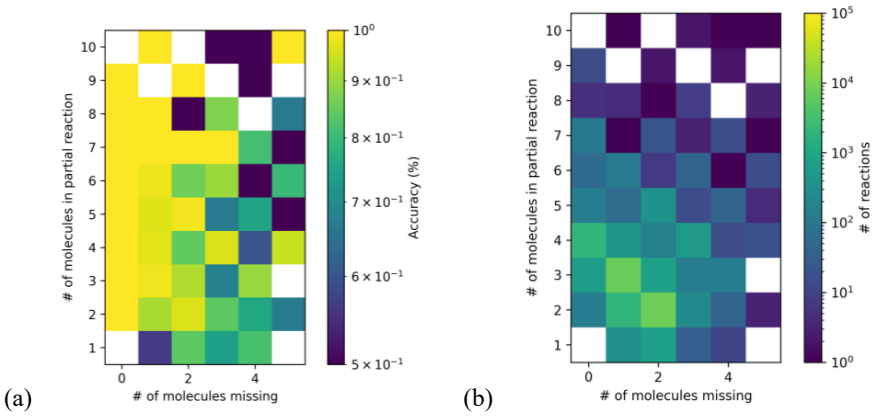

Figure 3. Model accuracy with degree of partialisation. In (a), we show the model accuracy across degrees of partialisation and in (b) we illustrate the corresponding amount of data across the degrees of partialisation.

3.2.2. Length of missing molecules

We also analyse the impact of the length of the missing molecules on the prediction accuracy and compare our results to reported results of the previously suggested encoder-only architecture, ChemMLM, (Zhang et al., 2023). Both models show a drop in accuracy when the length of the model output increases, see Table 3, yet the accuracy of the encoder-decoder architecture of this work decreases less. The autoregressive prediction of the encoder-decoder model takes previously predicted tokens into account, which helps the prediction of longer outputs, while the encoder-only RoBERTa architecture only considers non-masked information (incomplete reaction) for each mask prediction.

Table 3. Model accuracy rates per output length. Due to different tokenisation strategies, we translate our output length to their categories through the following: "short" corresponds to two atom tokens, "medium" to up to 20 atom tokens, and "long" to more than 21 atom tokens. ChemMLM results are not reimplemented, but taken from their work (Zhang et al., 2023).

Model type	“short” accuracy	“medium” accuracy	“long” accuracy
ChemMLM	99.9 %	78.3 %	16.4 %
Encoder-decoder	99.9 %	91.8 %	82.8 %

3.2.3. Language-based assumption without ground truth assumption

Considering the top five predictions, our LM predicts a reaction that is atom and charged balanced for 5.36 % of the reactions from the dataset without ground truth assumption. This is a drop from the high accuracy of the previous test set predictions based on the data with ground truth assumption and indicates larger differences between the data sets. Additionally, we tested the consistency of the model predictions as a basic sanity check through the round-trip accuracy. Here, at a partialisation where 70 % of the atoms are given, the model is relatively consistent in predicting the original reaction again (in 80 % of the cases considering top five).

4. Conclusions

In this work, we present a sequential hybrid approach for the completion of incomplete chemical reaction equations. We used a rule-based method to curate the bulk of incomplete reactions by applying mathematical rules based on atom and charge balances.

We fine-tuned the Molecular Transformer on the data set of complete reactions that we obtained from the rule-based completion algorithm. Our rule-based approach raised the completion rate from 3,5 % to 49.37 or 55.57 % depending on the set of helper species. The LM predicted reactions where atoms and charges are balanced with high accuracy for reactions in the test set, yet only for 5.36 % of the reactions in the remaining dataset. Future investigations are needed to better understand the dataset impact. Our results overall indicate the suitability of combined rule-based and machine learning based curation approaches and provides a further step towards complete chemical reaction data.

References

A. Arun, Z. Guo, S. Sung, A.A. Lapkin, 2023, Reaction impurity prediction using a data mining approach, Chemistry-Methods, e202200062

H. Gao, T.J. Struble, C.W. Coley, Y. Wang, W.H. Green, K.F. Jensen, 2018, Using machine learning to predict suitable conditions for organic reactions, ACS central science, 4(11), 1465-1476

B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. Luu Nguyen, S. Ho, J. Sloane, P.Wender, V. Pande, Retrosynthetic reaction prediction using neural sequence-to-sequence models, ACS central science, 3(10), 1103-1113

P.M. Jacob, T. Lan, J. M. Goodman, A. A. Lapkin, 2017a, A possible extension to the RInChI as a means of providing machine readable process data, Journal of Cheminformatics, 9, 1-12

P.M. Jacob, P. Yamin, C. Perez-Storey, M. Hopgood, A. A. Lapkin, 2017b, Towards automation of chemical process route selection based on data mining, Green Chemistry, 19(1), 140-152

D.M. Lowe, 2012, Extraction of chemical structures and reactions from the literature. Diss. University of Cambridge

P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C.A. Hunter, C. Bekas, A.A. Lee, 2019, Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction, ACS central science, 5(9), 1572-1583

P. Schwaller, A.C. Vaucher, T. Laino, J.L. Reymond, 2021, Prediction of chemical reaction yields using deep learning, Machine learning: science and technology, 2(1), 015016

K. Ulonska, M. Skiborowski, A. Mitsos, J. Viell, 2016, Early-stage evaluation of biorefinery processing pathways using process network flux analysis, AIChE Journal, 62(9), 3096-3108

A.Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, 2017, Attention is all you need, Advances in neural information processing systems 30

A.C. Vaucher, P. Schwaller, T. Laino, 2020, Completion of partial reaction equations, Chemrxiv

J.M. Weber, Z. Guo, C. Zhang, A.M. Schweidtmann, A.A. Lapkin, 2021. Chemical data intelligence for sustainable chemistry, Chemical Society Reviews, 50(21), 12013-12036

J.M. Weber, Z. Guo, A.A. Lapkin, 2022, Discovering Circular Process Solutions through Automated Reaction Network Optimization, ACS Engineering Au, 2(4), 333-349

L. Yinhan, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, 2019, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692

C. Zhang, A. Arun, A.A. Lapkin, 2023, Completing and balancing database excerpted chemical reactions with a hybrid mechanistic-machine learning approach. Chemrxiv