Completing and balancing database excerpted chemical reactions with a hybrid mechanistic - machine learning approach

Chonghuan Zhang,[1] *Adarsh Arun*[1,2,3] *and Alexei A. Lapkin*[1,2,3] *

[1] Department of Chemical Engineering and Biotechnology, University of Cambridge, Philippa Fawcett Drive, Cambridge CB3 0AS, United Kingdom

[2] Cambridge Centre for Advanced Research and Education in Singapore, CARES Ltd, 1 CREATE Way, CREATE Tower #05-05, 138602 Singapore

[3] Chemical Data Intelligence (CDI) Pte Ltd, 80 Robinson Road, #02-00, 068898, Singapore

∗ Corresponding author E-mail addresses: aal35@cam.ac.uk

https://doi.org/10.26434/chemrxiv-2023-hrgfw **ORCID:** https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

Abstract

Computer Aided Synthesis Planning (CASP) development of reaction routes requires understanding of complete reaction structures. However, most reactions in the current databases are missing reaction co-participants. Although reaction prediction and atom mapping tools can predict major reaction participants and trace atom rearrangements in reactions, they fail to identify the missing molecules to complete reactions. This is because these approaches are data-driven models trained on the current reaction databases which comprise of incomplete reactions. In this work, a workflow was developed to tackle the reaction completion challenge. This includes a heuristic-based method to identify the balanced reactions from reaction databases and complete some imbalanced reactions by adding candidate molecules. A machine learning masked language model (MLM) was trained to learn from reaction SMILES sentences of these completed reactions. The model predicted missing molecules for the incomplete reactions; a workflow analogous to predicting missing words in sentences. The model is promising for prediction of small and middle size missing molecules in incomplete reaction records. The workflow combining both the heuristic and the machine learning methods completed more than half of the entire reaction space.

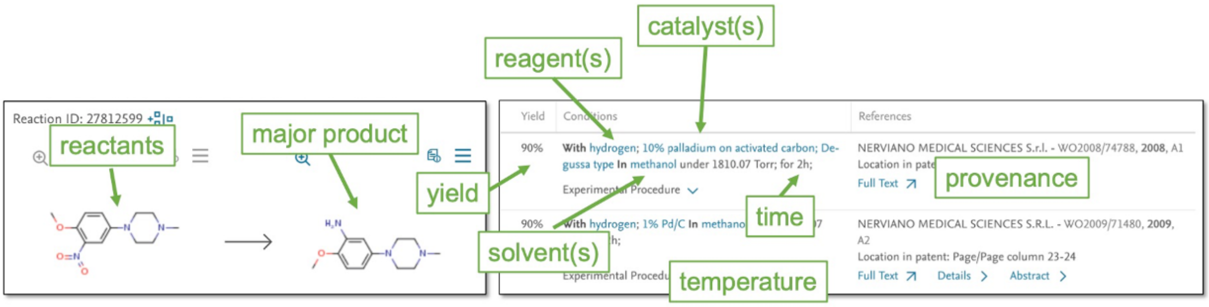
***Keywords**:* chemoinformatics; reaction informatics; organic synthesis; machine learning

Introduction

To enable evaluation of reaction routes with respect of a set of target parameters, such as overall yield, impurities, economy or greenness, knowledge of the complete reaction record is required. When all reaction participants (reactants, reagents and products) are known, the reaction completion is simply a problem of mass conservation, *i.e.* using linear algebra to balance the reaction with stoichiometric coefficients. However, this is not always the case for data records

https://doi.org/10.26434/chemrxiv-2023-hrgfw **ORCID:** https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

that are currently accessible in Reaxys®,[1,*] USPTO,[2] or any other none-manually curated reactions databases. An example of Reaxys® reaction data record is shown in Figure 1. Three main aspects of the reaction data are often missing in records today:[3] stoichiometric coefficients, reaction co-participants and integration of multiple reaction steps into a single reaction entry.



Reaction ID: 27812599

Yield	Conditions	References
90%	With hydrogen; 10% palladium on activated carbon; Degussa type In methanol under 1810.07 Torr; for 2h; Experimental Procedure	NERVIANO MEDICAL SCIENCES S.r.l. - WO2008/74788, 2008, A1 Location in pate Full Text
90%	With hydrogen; 1% Pd/C In methano 7 h; Experimental Procedu	NERVIANO MEDICAL SCIENCES S.R.L. - WO2009/71480, 2009, A2 Location in patent: Page/Page column 23-24 Full Text Details Abstract

Figure 1. An example reaction record (Reaxys® reaction ID: 27812599) summarised from multiple literature sources. The reaction record shows only major reactants and products, while the reaction co-participants are missing. The reaction participants, which do not contribute to the carbon flow, may be recorded as reagents.

There are historical and habitual reasons for incompleteness of reaction data today. Firstly, chemists report reactions in journal articles and patents based on the self-defined scope of research, which does not typically have potential tasks of others in mind; side products are either not in scope of studies, or were not detected by the analytical techniques that were used. Secondly, although text-mining techniques could process chemical information from literature, including properties and structures of molecules, and reaction conditions, sometimes it is hard to identify reaction participants as they may appear in different sections of a publication. This aspect is being addressed by creating clear templates for presenting and storing reaction data, such as the Open Reaction Database project.[4] For data to be useful for machine learning tasks

[*] *Copyright © 2022 Elsevier Limited except certain content provided by third parties. Reaxys is a trademark of Elsevier Limited*

https://doi.org/10.26434/chemrxiv-2023-hrgfw **ORCID:** https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

and for automated tasks of process development, it is necessary for existing data to be re-calibrated to include the missing reaction participants and reduce the noise in the datasets.

In literature, several methods for reaction structure completion were published. Grzybowski *et al.* manually curated around 100,000 generalised reaction rules with complete understanding of reaction participants and stoichiometry.[5] Their templates are now linked with commercial software SYNTHIA^{TM} and can guide retrosynthesis and analyse carbon efficiency based on mass conservation.[6] However, human development of reaction rules is far from the aspiration of exploring a very large chemical space of feasible molecules and reactions. On the other hand, the automatically extracted templates[7] are also not reliable for reaction completion, since most of these templates were generalised from open-source USPTO reaction dataset[2] in which the reactions were not complete. A “golden dataset” with complete reactions could be an ultimate solution. However, such large-scale reaction dataset does not exist today.

Atom mapping, which relies on rearrangement of atoms in chemical transformations, is promising to tackle this problem. Atom mapping describes the exact transformation and reveals the missing species on the product side. Jaworski *et al*. utilised graph-theoretical considerations and chose 20 chemical rules / heuristics to correct mapping of reactions.[8] This method attempts to complete stoichiometry firstly by adding small molecules such as acetaldehyde, ammonia, and others to balance the reactions and, secondly, by fitting reactions into popular reaction templates and adding the missing parts. Only if such attempts fail, atom mapping is employed.

The atom mapping tool RXNMapper developed by Schwaller *et al*. utilised NLP to infer reaction structures.[9] A neural network (transformer) was trained on a set of mapped reactions and proved to be capable of completing the mapping tasks quicker and with confidence scores.[9]

https://doi.org/10.26434/chemrxiv-2023-hrgfw **ORCID:** https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

Nugmanov *et al.*[10] developed a rule-based reaction balancing method and this has been integrated into a reaction informatics software, namely, CGRtools.[10] However, this could only add small molecules such as water, and the author claimed the balancing was imperfect. Thus, till now inferring complete reaction structure from incomplete reaction datasets remains a challenge.

In this work, a workflow was proposed towards reaction completion of existing reaction data records. Given a reaction record, either from reaction SMILES, or other reaction representations, a proposed heuristic tool, ChemBalancer, first checks if the reaction is a complete reaction, or incomplete from the left hand side (LHS) or the right hand side (RHS) of the reaction equation. The ChemBalancer intends to add one (or sometimes more than one) specie into a side (or sometimes two sides) of reaction accordingly, with the aid of reaction atom mapping tool RXNMapper.[9] If the ChemBalancer fails to complete the reaction record, the reaction SMILES string is considered as a language of chemistry and is passed into a fine-tuned masked language model (MLM), a BERT transformer,[11] originally designed to detect missing words in a sentence based on its context, to infer the most possible missing molecule in the reaction. Here, the MLM model, ChemMLM, was trained from a set of completed reactions detected by ChemBalancer, to predict the missing molecules in a reaction. This is followed by using ChemBalancer to further check if the updated reaction is completed. Every step of this pipeline intends to only add one most possible molecule to the incomplete reaction, and the specific sequential design of this pipeline was to learn only from available complete reactions and maximise the likelihood to propose correct molecules for incomplete reactions towards reaction structure completion.

https://doi.org/10.26434/chemrxiv-2023-hrgfw **ORCID:** https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

Methods

The heuristic step – ChemBalancer

ChemBalancer was partially adapted from Arun *et al.*'s balancing algorithm,[12] which was originally proposed as a step to identify chemical impurities produced in reactions. The workflow of ChemBalancer is shown in Figure 2, and summarised below.

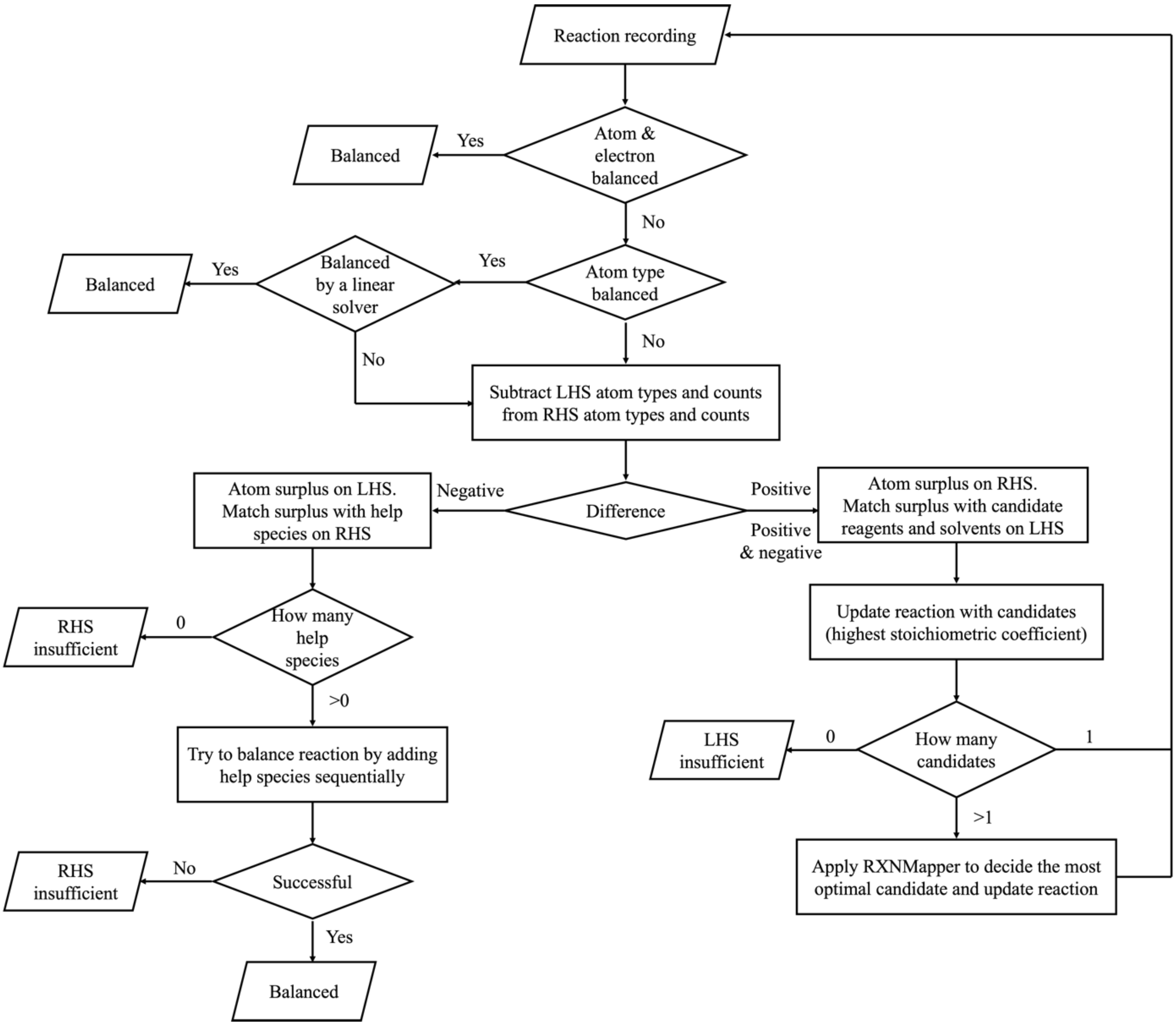

Reaction recording
Atom & electron balanced
Yes
Balanced
No
Atom type balanced
Yes
Balanced by a linear solver
Yes
Balanced
No
No
Subtract LHS atom types and counts from RHS atom types and counts
Difference
Negative
Atom surplus on LHS. Match surplus with help species on RHS
Positive
Positive & negative
Atom surplus on RHS. Match surplus with candidate reagents and solvents on LHS
How many help species
0
RHS insufficient
>0
Try to balance reaction by adding help species sequentially
Successful
No
RHS insufficient
Yes
Balanced
Update reaction with candidates (highest stoichiometric coefficient)
How many candidates
0
LHS insufficient
1
>1
Apply RXNMapper to decide the most optimal candidate and update reaction

Figure 2. A scheme of ChemBalancer workflow to detect if reaction records are originally complete, or could be completed with help species added to the RHS or candidate species added to the LHS, or the reactions are eventually incomplete with LHS or RHS species insufficient.

https://doi.org/10.26434/chemrxiv-2023-hrgfw ORCID: https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0

A reaction is defined as complete if atoms of each element and electron charges are balanced between LHS and RHS of a reaction equation. Hence, for a given reaction record, the first step for ChemBalancer is to determine the atom and the electron balances. By default, ChemBalancer ignores the counts of hydrogen atoms, since from perspective of solving linear algebra of reaction balancing, a valid reaction with mass conservation and all other atoms and electron charges balanced between LHS and RHS cannot give a degree of freedom for hydrogen imbalance. Hydrogen atoms should have been already balanced between LHS and RHS. If a reaction imbalance includes atom and electron counts, but the types (chemical elements) are identical, this reaction can be balanced with stoichiometry. A linear solver in ChemPy API[13] was used to determine stoichiometric coefficients. In some rare cases, large values of stoichiometric coefficients are added to reaction participants. Since the chance of having large numbers of molecules react together to trigger a reaction is low, an arbitrary value of six was set as an upper limit for stoichiometric coefficients. Here we must differentiate the small molecules synthesis from polymerisation reactions which could be represented by chemical equations with very large stoichiometric coefficients.

If stoichiometry alone cannot solve reaction balance, ChemBalancer attempts to add possible missing molecules to complete the reaction of interest. To do this, ChemBalancer subtracts the counts of each atom and electrons at the LHS from the RHS.

If this results in atom surplus on the LHS, one compatible small help specie is added to the RHS. A library of manually curated help species was tailored to include 32 molecules and ions, ranging from water to chlorobenzene. These molecules are the most frequently appearing side products in organic synthesis. The full list of help compounds is shown in Table 1. These help species are added sequentially to the RHS of a reaction and used to balance the reaction. The

https://doi.org/10.26434/chemrxiv-2023-hrgfw ORCID: https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0

reaction is defined as complete with help species if it is balanced with the addition of these species to the RHS. Otherwise, it is classified as a 'RHS species insufficient' reaction.

Table 1. The library of help compounds.

Categories	**Molecules**
Ion	sodium ion, potassium ion, nitronium ion, sulphate ion
Oxyacid	phosphoric acid, sulfuric acid, nitric acid, chloric acid
Alcohol	methanol, ethanol, propanol, butanol
Carboxylic acid	acetic acid, propanoic acid, butanoic acid, methoxyacetic acid
Aromatic compound	methane, benzene, toluene, phenol, chlorobenzene
Others	water, hydrogen, oxygen, ammonia, nitrogen
Hydrogen acid	hydrogen chloride, hydrogen bromide, hydrogen iodide, hydrogen fluoride, phosphine, hydrogen sulfide

If this step presents an atom surplus on the RHS, this means that at least one reactant is missing for the reaction. The most likely reactants that may complete reactions are listed as one of reagents or solvents in the reaction record. For example, in Reaxys®, reagents, *i.e.* reactants that do not contribute to carbon flow of the reaction) and solvents are listed as two attributes in reaction record, as shown in Figure 1. In USPTO, reactions are given with reaction SMILES strings in the format of "A.B>C.D>E.F", in which "C" and "D" are reagents, whilst solvents are not provided. ChemBalancer picks a molecule from the candidate reagents and solvents to add to the LHS of the reaction. If the reaction has multiple candidate reagents and solvents, the atom mapping tool RXNMapper[9] is implemented to map the atom rearrangement from the reactants to the products for each candidate reaction. RXNMapper can trace the reactant' origin of each atom of the product. Given a reaction with RHS atom surplus this helps identify the percentage of product atoms that can be traced with their origins at the LHS, and this is

https://doi.org/10.26434/chemrxiv-2023-hrgfw ORCID: https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0

quantified by the atom mapping confidence score.[9] The confidence scores for the candidate reactions are ranked to select the most optimal candidate molecule to update the LHS of the reaction. ChemBalancer treats the updated reaction as a new reaction record and tries to balance the updated reaction again with the same workflow. This corresponds to the arrow from the bottom-right box to the top box in Figure 2.

Each loop of ChemBalancer aims to add one possible missing molecule to the reaction to complete it. However, in case when a candidate reagent or solvent is added to the LHS, the reaction reaches a new imbalance status. If it is passed into the workflow in a new balancing loop, this may potentially result in adding more than one molecule to both sides of the reaction. The reaction is eventually defined as 'LHS insufficient' reaction when there are no available reagents or solvents to be added to the LHS to balance it.

If a reaction has atom surplus on the LHS for some atom types and atom surplus on the RHS for others, this means the reaction is missing both - at least one reactant and one product. With given candidate reagents and solvents, the reaction completion starting from adding species into the LHS becomes more certain and has higher priority than the RHS. Therefore, this case follows the same procedure as the reaction with atom surplus on the RHS. However, since atom surplus exists in both sides of the reaction, the possibility of adding only candidate reagents and solvents on the LHS to complete the reactions becomes very low. Therefore, help species are also allowed to be added to the LHS as well, but with a lower priority than the candidate reagents and solvents, since they are prone to create false positive complete reactions.

https://doi.org/10.26434/chemrxiv-2023-hrgfw ORCID: https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0

The machine learning step – ChemMLM

Those reactions that were failed to be balanced by ChemBalancer were marked as incomplete reactions and passed to ChemMLM to infer the next most possible molecules in the LHS or the RHS of reaction record, based on conclusions with regards of LHS or RHS- insufficient species reached by ChemBalancer.

Several assumptions were made in developing ChemMLM:

1. All results passed from ChemBalancer were assumed to be correct. While the originally balanced reactions verified by ChemBalancer are usually correct, there remain false positives in the reactions completed by ChemBalancer itself. Also, since ChemMLM was trained on a set of completed reactions detected by ChemBalancer, all reactions in this set were assumed correctly completed by ChemBalancer. However, in reality, false positives are present in this set and they would result in propagated error in training and reduce ChemMLM accuracy.
2. ChemMLM assumes only one molecule is missing in incomplete reactions. Adding more than one molecule would result in combinatorial increase of decision space, which is not guaranteed to lead to ground truth molecules but becomes computationally expensive. In this case, ChemMLM was trained to predict the most possible next molecule to be added to the reaction towards reaction completion. However, we noticed that sometimes ChemMLM predicts more than one molecule in SMILES format, using representation "A.B". This is discussed in the following sections.

The ChemMLM model structure

MLM is an application of BERT transformer.[11] Using BERT to learn from many linguistic patterns of complete sentences, MLM predicts missing words in a sentence inferred from its

https://doi.org/10.26434/chemrxiv-2023-hrgfw ORCID: https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0

context. For example, to fill the gap "*In autumn, the* ___ *falls from the trees*", MLM restricts search to the most possible item falling from trees in autumn, which would be the word "*leaves*". To train an MLM, a BERT transformer is fed with the same input sequence as the output, and the model optimises weights within its encoder layers in order to process this. To predict a missing word, tokens are randomly masked within the input sequence by replacing the missing word token with a special mask token symbol "<mask>". Given the context of a sentence, the MLM learned from previous semantic examples would infer the possible missing word.[14]

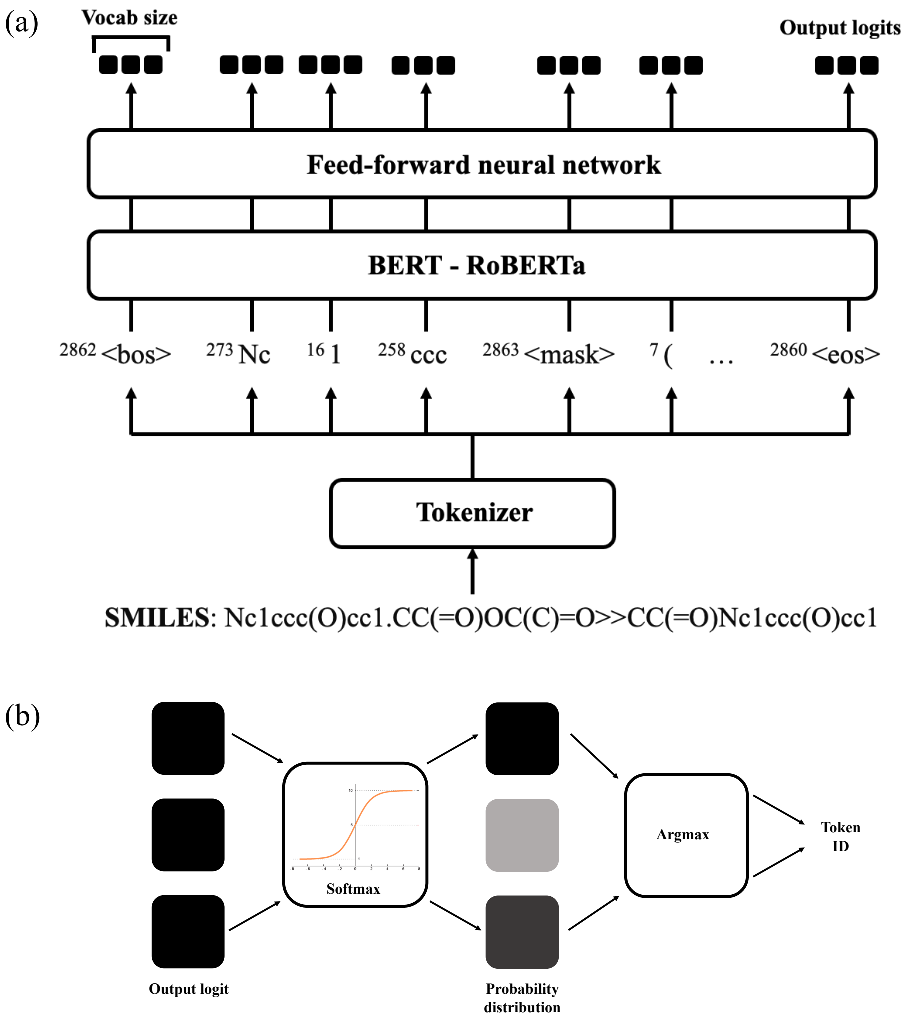


Figure 3. A diagramme of model structure of ChemMLM: (a) transformation of reaction SMILES strings into output logits by passing the tokeniser and RoBERTa model,[15] using an example of reaction SMILES of paracetamol synthesis, acetylation of 4-aminophenol with

https://doi.org/10.26434/chemrxiv-2023-hrgfw ORCID: https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0

acetic anhydride; and (b) transformation of output logits into token IDs by passing softmax and argmax functions.

Figure 3 shows a proposed model structure for ChemMLM adapted from an MLM model.[14] In Figure 3(a), reaction SMILES strings are processed into ChemMLM. A reaction SMILES is tokenised into multiple tokens and replaced by index IDs that have one-to-one correspondence with the tokens. The reaction SMILES tokenisation method is discussed in "ChemMLM tokeniser" Section of this paper. A mask function dynamically masks a section of reaction SMILES, where the mask method is discussed in the "Mask method" Section of this paper.

The input ID tensor is processed by a BERT transformer model, where the RoBERTa[15] variant of BERT is adopted. The theories and model architecture of RoBERTa are discussed in the "RoBERTa model" Section of this paper. RoBERTa outputs a set of vectors with the length of 768, and each vector is transformed from a token by RoBERTa. The vectors are passed into a feed-forward neural network, which outputs the output logits. The output logit is the logit transformation applied to its original output, shown in Eq. *1*. The output vector of each token remains the size equal to the vocal size, *i.e.* the total number of tokens, indicating the projection of each tokens into each output logit.

$$logit(output) = \frac{output}{1 - output} \qquad \text{Eq. 1}$$

The step to convert each set of output logits into an output token is shown Figure 3(b). From the output logits, a softmax transformation is applied to acquire probability distribution of a list of possible predicted tokens. This is followed by applying an argmax function to select the most possible index ID for the token. Concatenating the strings of tokens produces a predicted reaction SMILES, which reveals the masked tokens represented in the mask symbol "<mask>".

https://doi.org/10.26434/chemrxiv-2023-hrgfw **ORCID:** https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

The input tokens are compared with the predicted tokens through Kullback-Leibler (KL) divergence loss function, shown in Eq. 2, which pairwise measure the probability distribution of the true output encoding y_{true} and computed one y_{pred}. The loss is backpropagated to the encoder layers of RoBERTa, the feed-forward neural network layers, and the softmax function to update the weights.

$$KL\left(y_{pred}, y_{true}\right) = y_{true} \cdot \log \frac{y_{true}}{y_{pred}} \qquad \text{Eq. 2}$$

The ChemMLM tokeniser

Ideally, in an MLM model, each word is converted into a single token, and prediction of a missing word is prediction of one token. However, this could not be the case for ChemMLM. With 160 million molecules present in the database, conversion of a single molecule into a token would result in a vocabulary size of 160 million, which was impossible to process in ChemMLM.

A well-known regularised reaction SMILES tokenisation method has been proposed by Schwaller *et al.*[16], to deliver the promising reaction prediction tool - Molecular Transformer. In this method, all atoms and regular used expressions in reaction SMILES are separated into tokens, such as "C", "Br", "@" and ")". Although this regularised tokenisation was implemented in multiple related works, it was not used here. This approache discretises a molecule into multiple tokens, which makes it harder to learn from the semantic context of a molecule's SMILES string.

A byte-pair encoding (BPE) tokenisation method[17, 18] was used in this work. The BPE tokenisation counts the highest frequent consecutive string expressions of reaction SMILES to

https://doi.org/10.26434/chemrxiv-2023-hrgfw ORCID: https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0

replace with tokens. For example, in Figure 3(a), “Nc” and “ccc” are both high frequent expressions in reaction SMILES, and therefore, they were formed as single tokens. This tokenisation method disconnected molecules at various positions, which allowed the model to learn different molecular disconnection strategies, and this also steered the model to infer various types of missing molecules.

High frequent expressions were usually present within the molecule representations of the reaction SMILES, *i.e.* the splitting molecule symbol “.” and the splitting reactant-product symbol “>>” do not exist in the tokens. However, in several cases, tokens were formed across the molecule representations, since the BPE tokenisation method recognises SMILES strings of “part of a molecule’s SMILES + . + part of another molecule’s SMILES” as high frequent expressions, such as “+].[” and “.[”.

Four special symbols were added into the token vocabulary, which are “<bos>”, “<eos>”, “<mask>” and “<pad>”, meaning respectively, the start token of a reaction SMILES string, end token of a reaction SMILES string, the mask token and placeholder of void tokens. Using all Reaxys® and USPTO reactions as semantic input, the BPE method tokenised the reaction SMILES expression into 2,863 tokens including the four special tokens.

RoBERTa model

RoBERTa model[15] builds on the BERT transformer,[11] and modifies key hyperparameters of the BERT transformer. The choice of RoBERTa over the original BERT here was because of its use of BPE tokenisation approach, as discussed in the “ChemMLM tokeniser” Section, and dynamic mask method, as discussed in the “Mask method” Section. The other architecture of RoBERTa is close to BERT.

https://doi.org/10.26434/chemrxiv-2023-hrgfw **ORCID:** https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

RoBERTa, or BERT only use the encoding half of the transformer, as the decoding half is not required to translate the input sequence into another language.[11, 15] In Figure 3, the "BERT-RoBERTa" unit includes the input embedding unit, positional encoding unit, multi-head attention units, and normalisation layer, followed by layers of feed-forward neural network. These units enable the ChemMLM to encode the semantic and syntactic information of the reaction SMILES in the embeddings.

ChemMLM model training and prediction

Data Source

Since a "golden dataset" with a large range of complete reactions did not exist at the start of this work, all accessible complete reactions were valuable data sources, despite doubts about data quality of these complete reactions. The objective was to explore the limited complete reaction space, to broaden the boundary of the space to some originally incomplete reactions, while retaining errors and noise present in the space. Therefore, the complete reactions, *i.e.* the originally complete reactions detected by ChemBalancer and the originally incomplete reactions balanced by ChemBalancer, were used to train the ChemMLM model. Reaxys® and USPTO reaction datasets were two accessible reaction data sources. Although the data size of the two dataset are in different magnitude - approximately 21 million versus one million, they both cover broad ranges of reaction types,[19] which are both key data sources to learn from.

To compare the model predictability from different data sources, ChemMLM models were trained from the complete reactions from (i) USPTO dataset and (ii) Reaxys® plus USPTO dataset (this is referred to as the combined dataset in the rest of this paper). Since the magnitude

https://doi.org/10.26434/chemrxiv-2023-hrgfw **ORCID:** https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

of USPTO is not comparable with Reaxys, it was added into Reaxys® as a whole data source to train the second ChemMLM model.

The data was split into training, validation and test data following the ratio of 9 : 0.5 : 0.5, to train, evaluate the potential model structure, and understand the model predictability respectively. Only a small portion of reactions were used for model validation and test, since the total number of reactions is large.

Mask method

A dynamic mask method randomly masked approximately 15% of the tokens in the input ID tensor. The number of masked tokens was rounded based on the number of tokens in the reaction SMILES. To avoid using the same masks for every epoch of training, the method randomly masked the reaction SMILESs iteratively in every epoch, to increase the exposure of every token as a mask.

Model implementation

The ChemMLM models were implemented with a Python API, namely Hugging Face,[20] which is a platform to implement multiple NLP transformer variants. The training calculations of the ChemMLM models were powered by Google Colab's GPU cloud service.[21]

The model structure details configured for the RoBERTa model are shown below. The dimensionality of the encoder layers is 768. The maximum token length that the model can be used with is 512, which covers all reaction SMILES token length (up to 333). In the encoder, the number of attention heads for each attention layer is 12, and the number of hidden layers is 6. The activation function used in the model is Gaussian Error Linear Units (GELU) function.[22]

https://doi.org/10.26434/chemrxiv-2023-hrgfw **ORCID:** https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

The dropout probability for all fully connected layers in the embeddings and encoder is 0.1, and the dropout ratio for the attention probabilities is 0.1. The epsilon used by the normalisation layers is 10^{-12}. The configuration of such a model creates 83,504,416 parameters.

After fine-tuning of ChemMLM training arguments based on the validation data results, the model arguments for the final ChemMLM models trained from USPTO data, and combined data are shown in Table 2, and “adam” optimiser was used to optimise the model parameters.

Table 2. Ainal training arguments for two ChemMLM models, learned from USPTO data and combined data respectively.

ChemMLM	USPTO	Combined
Epoch	175	17
Learning rate	10^{-4}	1.5×10^{-4}
Gradient accumulation steps	1	1
Batch size	16	32

Model test method

The trained ChemMLM models were assessed on the reaction test dataset. The model used KL divergence loss, determined in Eq. 2 between predicted tokens and the true masked tokens to backpropagate the model parameters and validate the models. However, this assessment could only conclude on the model predictability of a masked token, which is usually part of a molecule. The purpose for the model testing was to assess whether the model could infer an entire missing molecule, rather than a masked token. Moreover, since the model predicts the missing molecule for the incomplete reactions following the LHS and RHS species insufficient scenarios, as discussed in the “Missing molecule prediction” section, the ChemMLM models were tested, with respect to its LHS and RHS predictability.

https://doi.org/10.26434/chemrxiv-2023-hrgfw ORCID: https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0

With respect to model predictability testing at the RHS, in each complete reaction SMILES string in the test dataset, each product was hidden once alternatively with the missing molecule symbol “@@@”. For example, a reaction with format of “A.B.C>>D.E” would be duplicated into two reaction SMILESs, “A.B.C>>@@@.E” and “A.B.C>>D.@@@”. With respect to the LHS testing, in each reaction, each reactant was hidden once alternatively. The example reaction would be duplicated into three reaction SMILESs, “@@@.B.C>>D.E”, “A.@@@.C>>D.E” and “A.B.@@@>>D.E”. In this way, the number of reactions in the test dataset was also augmented, and the LHS ChemMLM and the RHS ChemMLM were used to predict the hidden molecules at each side respectively.

Each “@@@” symbol was converted to the multiple of the mask symbol “<mask>” based on the number of tokens included in the hidden molecule. In this way, instead of dynamically randomisng 15% of the mask, the masks in the test dataset were tailored to mask an entire molecule sequentially at the designated side of the reaction. The ChemMLM model predicted the masked tokens and concatenated the tokens to predict SMILES of the hidden molecule.

Any deviation in tokens between the predicted molecule and the ground truth would make prediction of the molecule semantically meaningless. Therefore, instead of comparing the KL divergence loss, as used in model training and validation, SMILESs of the predicted molecules and the ground truth molecules are compared directly to determine the correction rate of identical SMILES strings among the test dataset. For each ChemMLM model, two values were computed, *i.e.* the LHS and the RHS correction rates in the test dataset.

The hidden molecule mask method was initially also considered as a regularised mask method to mask molecules during training instead of the dynamic mask approach. However, such a

https://doi.org/10.26434/chemrxiv-2023-hrgfw ORCID: https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0

mask method hides a molecule at a centralised segment (*i.e.* one single token) of the reaction SMILES. By no means could the mask method learn bond breaks and recombination from the reaction SMILES, and it was hard to pick the semantic context among the molecules. Therefore, a dynamic mask approach was implemented.

Missing molecule prediction

Based on the suggestion from ChemBalancer - a reaction is imbalanced with either LHS or RHS species insufficient. Assuming only one molecule is missing in the reaction SMILES, an extra missing molecule symbol “@@@” of a missing molecule is added to the corresponding side of the reaction SMILES. For the paracetamol synthesis discussed above, ChemBalancer suggested RHS species insufficient for the reaction, and therefore, “@@@” symbol was added to the end of the reaction SMILES RHS, split by splitting molecule symbol “.” from the original reaction SMILES, shown as “Nc1ccc(O)cc1.CC(=O)OC(C)=O>>CC(=O)Nc1ccc(O)cc1.@@@”. This example is shown in Figure 4.

https://doi.org/10.26434/chemrxiv-2023-hrgfw **ORCID:** https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

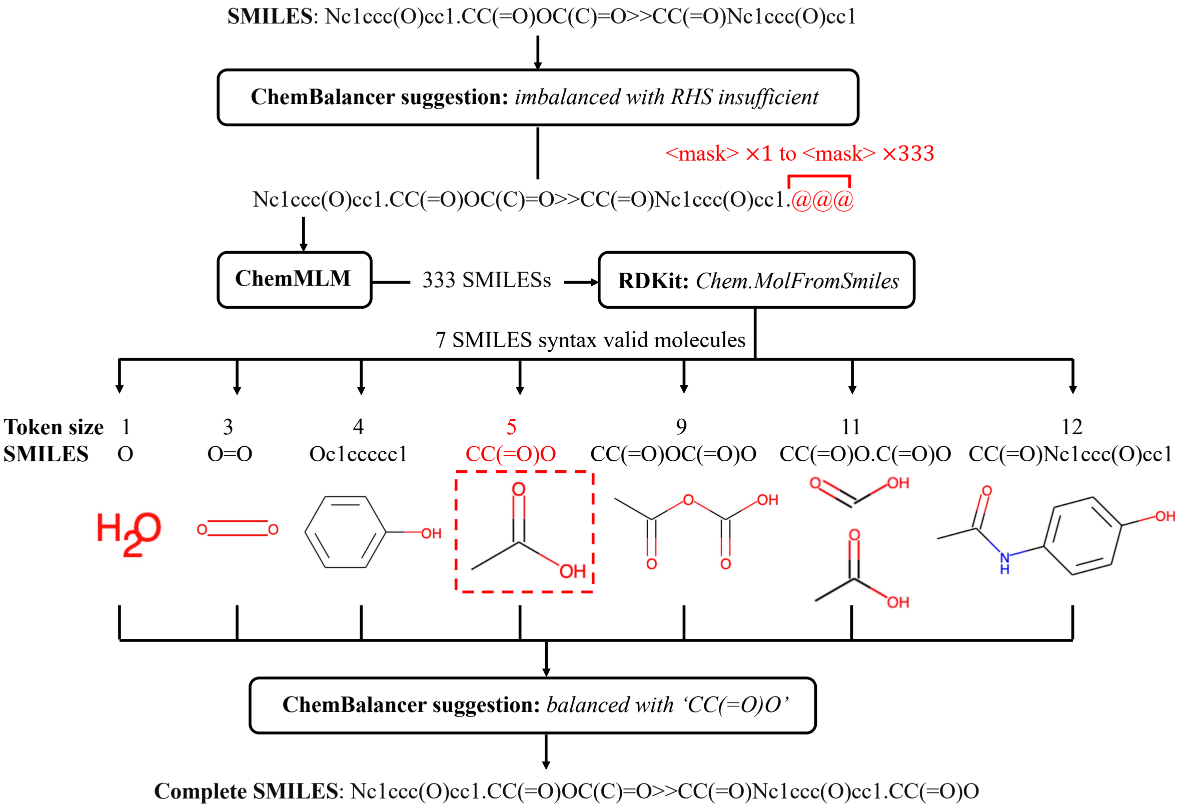

SMILES: Nc1ccc(O)cc1.CC(=O)OC(C)=O>>CC(=O)Nc1ccc(O)cc1
ChemBalancer suggestion: imbalanced with RHS insufficient
<mask> ×1 to <mask> ×333
Nc1ccc(O)cc1.CC(=O)OC(C)=O>>CC(=O)Nc1ccc(O)cc1.@@@
ChemMLM
333 SMILESs
RDKit: Chem.MolFromSmiles
7 SMILES syntax valid molecules
Token size 1 3 4 5 9 11 12
SMILES O O=O Oc1ccccc1 CC(=O)O CC(=O)OC(=O)O CC(=O)O.C(=O)O CC(=O)Nc1ccc(O)cc1
H2O
O
OH
N
H
ChemBalancer suggestion: balanced with 'CC(=O)O'
Complete SMILES: Nc1ccc(O)cc1.CC(=O)OC(C)=O>>CC(=O)Nc1ccc(O)cc1.CC(=O)O

Figure 4. An illustration of the workflow to predict the missing molecule for an incomplete reaction, illustrated with an example of a RHS species insufficient reaction, paracetamol synthesis. The reaction is completed with acetic acid molecule added to RHS of the reaction.

The prediction of a known hidden molecule from a complete reaction is different from the prediction of a missing molecule from an incomplete reaction. While the number of tokens of the hidden molecule from the complete reaction is given, that of the missing molecule from an incomplete reaction remains unknown. A molecule in USPTO and Reaxys® reactions can have token lengths ranging from one to 333. The 333 scenarios are all enumerated, by converting the missing molecule symbol "@@@" in reaction SMILES into different numbers of mask symbols "<mask>". For example, for the scenario of the missing molecule with two tokens, the paracetamol synthesis reaction SMILES is converted to "Nc1ccc(O)cc1.CC(=O)OC(C)=O>>CC(=O)Nc1ccc(O)cc1.<mask><mask>".

https://doi.org/10.26434/chemrxiv-2023-hrgfw ORCID: https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0

ChemMLM was used to predict the masked tokens and concatenate them into a molecule. It is noticed that in some cases, more than one molecule was concatenated from the masked tokens, since the model can predict crossing-molecule token, which splits the concatenated reaction SMILES tokens into multiple molecules.

From the 333 prediction enumerations, only one token length could potentially give the ground truth for the missing molecule, and the other 332 enumerations are semantically meaningless predictions, and even sometimes, the SMILES syntax of the predictions is incorrect. To determine the meaningful result, RDKit API[23] was used to convert the 333 candidate molecular SMILES strings into molecule objects, and the syntax meaningless molecules would throw error and be removed from the candidate molecules. For example, for the paracetamol synthesis reaction, only 7 out of the 333 enumeration produce SMILES syntax valid molecules, which have the token lengths of 1, 3, 4, 5, 9, 11 and 12 respectively. Among these scenarios, the 11 tokens concatenate into two molecules, with "." presented in the SMILES string, as shown in Figure 4.

These syntax valid molecule SMILESs replace the "@@@" in the reaction SMILES and are passed to ChemBalancer to determine if these candidate molecules include a chemically meaningful molecule that could either complete the reaction itself, or could balance the reaction with extra molecules proposed by ChemBalancer. If both scenarios fail, this means the reaction could not be completed with the current workflow, as shown in Figure 5.

https://doi.org/10.26434/chemrxiv-2023-hrgfw ORCID: https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0

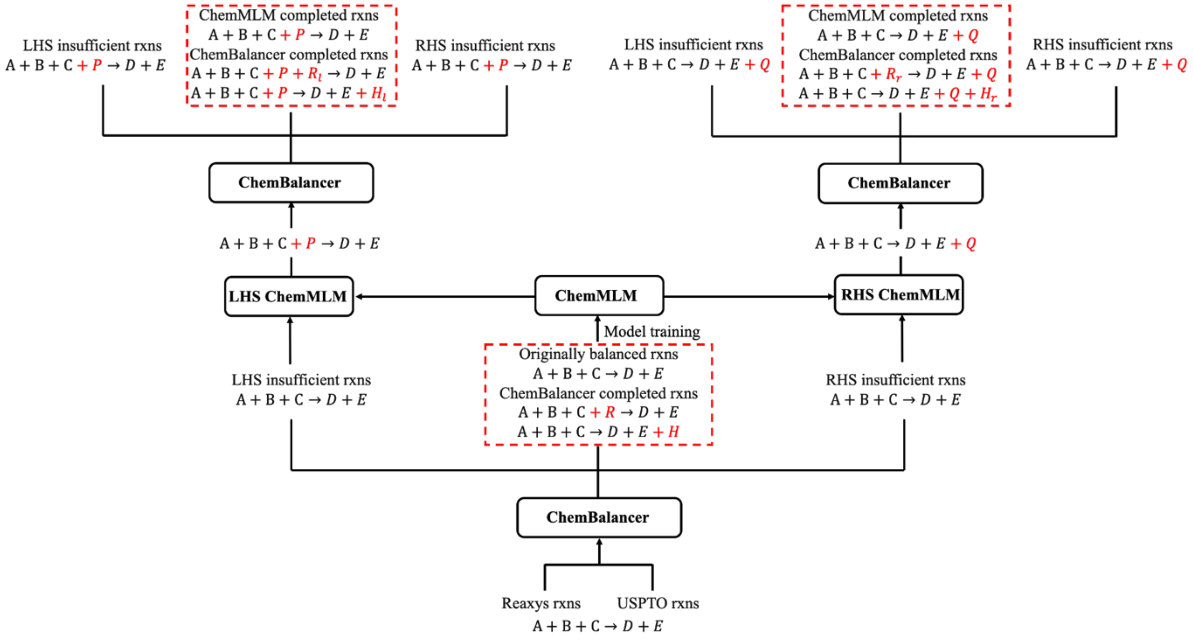


Figure 5. An illustration of the entire workflow of the reaction completion algorithm, including ChemBalancer, and ChemMLM subunits, exemplified by a reaction scheme “A+B+C → D+E”. All complete reactions are shown in red dashed boxes.

Reaction completion workflow

The entire workflow of the reaction completion algorithm includes two sequential subunits, ChemBalancer and ChemMLM. The workflow is summarised in Figure 5. In the ChemBalancer subunit, the USPTO and Reaxys® reactions were passed to ChemBalancer to determine if each reaction, exemplified by a reaction scheme “A+B+C → D+E”, is originally complete, or could be balanced with ChemBalancer by adding a candidate reagent “R” or a help compound “H” into the reaction, or sometimes combinations of “R” and “H". The complete reactions, shown in the bottom red dashed box of Figure 5 were used to train the ChemMLM model.

In the ChemMLM subunit, for the ChemBalancer-determined LHS insufficient reactions, the trained ChemMLM proposes multiple solutions for an extra reactant “P”. ChemBalancer then detects which proposed reaction “A+B+C+P → D+E” is balanced, or could be balanced by adding a reagent “R to the LHS or a help compound H_1 to the RHS, or sometimes combinations

https://doi.org/10.26434/chemrxiv-2023-hrgfw ORCID: https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0

of R_l and H_l. Similarly, for the RHS insufficient reactions, ChemMLM proposes multiple solutions for an extra product Q. ChemBalancer then detects which proposed reaction A+B+C → D+E+Q is balanced, or could be balanced by adding a reagent R_r or a help compound H_r, or sometimes combinations of R_r and H_r. When multiple completion solutions are proposed for a given reaction, the atom mapping tool RXNMapper selects the most optimal solution based on the highest ranked confidence score on atom mapping.

All complete reactions are shown inside the red dashed boxes in Figure 5. These reactions could potentially be used to train another round of the ChemMLM model for further broadening the complete reaction space. However, this was not conducted, since the workflow carries erroneous assumptions. These errors would be propagated into another round of training if these data were used to retrain the ChemMLM model, and further reduce the predictability.

Results and Discussion

The ChemBalancer completion results

ChemBalancer attempted to complete the reaction structures from the available resources (the candidate reagents and solvents, and the help species) by a heuristic approach. All reaction records from Reaxys® and USPTO databases were passed to ChemBalancer, with their completion results shown in Table 3. In USPTO and Reaxys®, 17.6% and 44.1% of reaction records were complete with ChemBalancer respectively. These complete reactions include the originally complete reactions, reaction completed by adding candidate reagents to the LHS (or, reagents to the LHS and help species to the RHS), and reaction completed by adding help species to the RHS. Table 3 also shows the ratio of reactions failed to be completed by ChemBalancer, with species insufficient at the LHS or RHS of the reactions. The reactions have both atom surplus in the LHS and RHS of the reactions were classified into the LHS species

https://doi.org/10.26434/chemrxiv-2023-hrgfw ORCID: https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0

insufficient reactions, since the priority to add species to the LHS was higher than RHS. The majority of the incomplete reactions are RHS species insufficient.

Table 3. Statistics for the reaction completion results by ChemBalancer for the reaction records in USPTO and Reaxys® databases.

Reaction database	**USPTO**	**Reaxys®**
Complete reaction	17.6%	44.1%
-- Originally complete reactions	3.3%	7.2%
-- Completed with reagents	0.5%	6.2%
-- Completed with help species	13.8%	30.7%
LHS species insufficient	2.1%	13.4%
RHS species insufficient	80.3%	42.5%
Total complete reactions	**171,637**	**7,043,030**

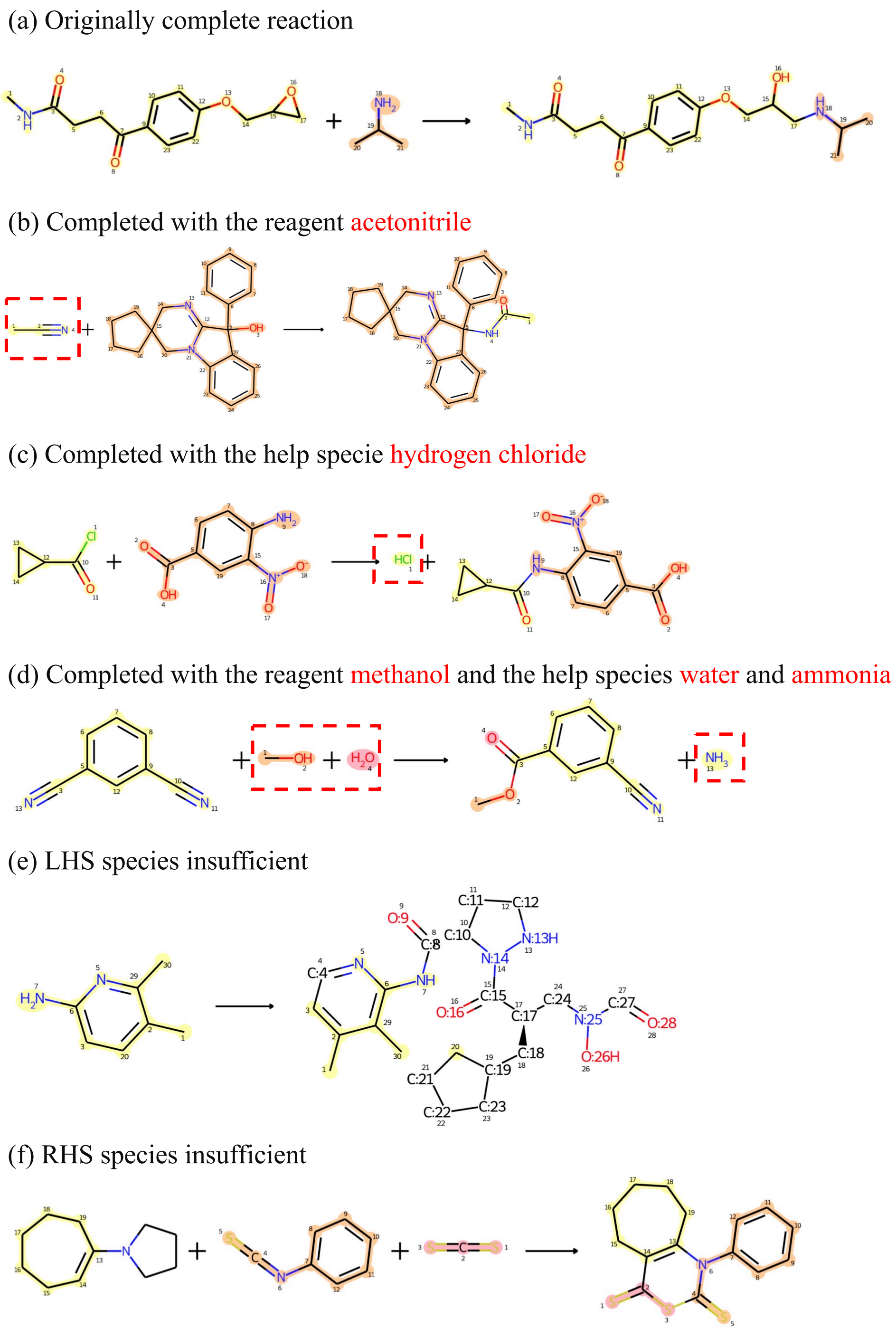
It is noticed that ChemBalancer has better performance on Reaxys® reactions over USPTO. A very high percentage of reactions in USPTO remains incomplete due to their RHS species insufficiency. This means that most reactions could not identify their side products from the library of help species. A larger set of help species was attempted to have more diverse solutions to complete the reactions. This list included 3,526 frequently appearing small species. However, from a set of LHS atom surplus reactions, enlarging the library of help species from 32 to 3,526 frequently appeared molecules as reaction side products was also attempted. However, this could only improve the reaction completion rate by 0.3%, but significantly increased computational time for the enumeration. This suggests that current 32 help species are sufficient to comprise the help species library. This also indicates a case-specific heuristic method, or a predictive machine learning is preferred to predict the exact side products for the RHS insufficient reactions. It is also noticed that the ratio of reactions with RHS species insufficiency is much higher in the USPTO reactions. This might be because the USPTO reaction extraction algorithm[24] incorrectly added reagents into the reaction SMILES as reactants. Using the atom mapping tool RXNMapper, it is noticed that the reactants from a great number of USPTO reactions have no atom mapped into the reaction products. With no contribution to the carbon

https://doi.org/10.26434/chemrxiv-2023-hrgfw **ORCID:** https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

flow, these reactant molecules were supposed to be added into the reagent categories. Adding these into reactants would cause LHS atom surplus, and therefore result in RHS species insufficiency. The USPTO complete reaction ratio would be higher if these false labelled reactants could be removed.

ChemBalancer eventually obtains 171,637 and 7,043,030 complete reactions respectively in USPTO and Reaxys® by removing the redundant reactions and reactions with molecules not sanitisable by RDKit. These reactions cover large reaction spaces for ChemMLM to learn from. Reaction examples for each completion and incompletion scenarios are shown in Figure 6. The ChemBalancer-added reagents and help species, shown in Figure 6(b-d) were evaluated from their atom rearrangement by the atom mapping tool RXNMapper,[9] and also manually by human experts from their reaction mechanisms, to be correct predictions for these example reactions.

https://doi.org/10.26434/chemrxiv-2023-hrgfw ORCID: https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0



(a) Originally complete reaction

(b) Completed with the reagent acetonitrile

(c) Completed with the help specie hydrogen chloride

(d) Completed with the reagent methanol and the help species water and ammonia

(e) LHS species insufficient

(f) RHS species insufficient

Figure 6. Examples of ChemBalancer reaction completion results: (a-d) complete reactions, (e) a LHS species insufficient reaction, and (f) a RHS insufficient reactions. The molecules added by ChemBalancer are shown in the red dashed boxes. All reactions are atom mapped by RXNMapper[9] and labelled with atom mapping indices.

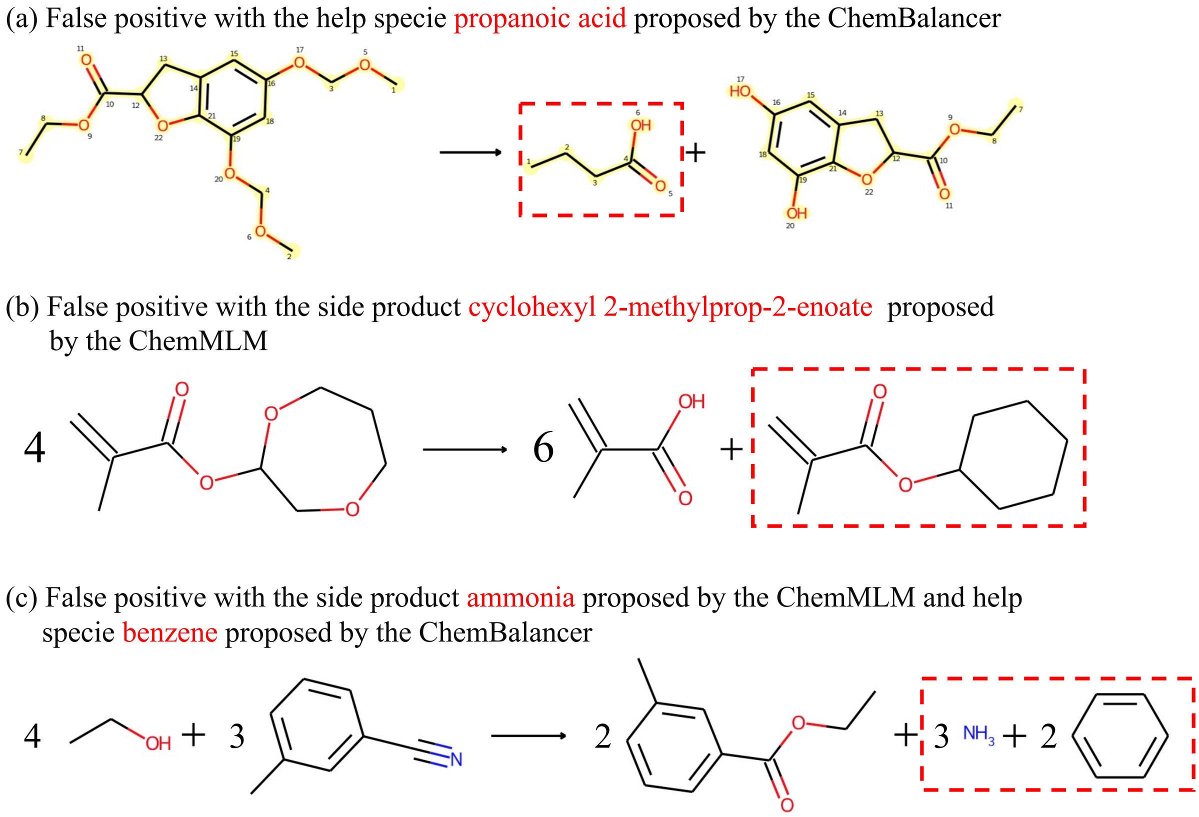
The reaction given in Figure 6(d) shows an example of a reaction completion by adding a reagent methanol and a help specie water at the LHS and adding a help specie ammonia at the RHS. The original reaction record, *i.e.* the reaction participants outside the red dashed boxes,

https://doi.org/10.26434/chemrxiv-2023-hrgfw **ORCID:** https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

has atom surplus of nitrogen at the LHS and atom surplus of carbon and oxygen at the RHS. ChemBalancer prioritised to first solve the RHS atom surplus, by adding one of the candidate reagents and solvents to the LHS. RXNMapper suggested that among the candidates, adding methanol would give the highest mapping confidence score. Afterwards, by adding help species of water at the LHS and ammonia at the RHS, the updated reaction was balanced. Help species were only allowed to be added to the LHS when the reaction has atom surplus at both sides. This reaction follows the mechanism of Pinner reaction, *i.e.* under an acidic environment, an alcohol esterificates a nitrile with water to form an imino ester.[25]

ChemBalancer balanced a large number of incomplete reactions by adding help species to the RHS of reactions. Although these help species were the most frequent species present as the side products of the reactions, by adding these help species as side products, these reactions were only completed in terms of material balance at the two sides of the reactions. However, the reaction mechanisms were not evaluated and, therefore, some help species completed reactions remain false positive. In contrast, the reagents or solvents balanced reactions were rare to be incorrect, since these candidate reagents and solvents are clues to complete the reactions given as reaction attributes. Examples of false positive reactions completed by ChemBalancer are shown in Figure 7(a). In this reaction, it is clear that the carbon-oxygen bonds (specifically with atom indices, the “C[:3]-O[:17]” and “C[:4]-O[:30]” bonds) of the reactant need to be disconnected by an reagent. However, ChemBalancer only proposes a false positive product propanoic acid to the reaction.

https://doi.org/10.26434/chemrxiv-2023-hrgfw **ORCID:** https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0


(a) False positive with the help specie propanoic acid proposed by the ChemBalancer

(b) False positive with the side product cyclohexyl 2-methylprop-2-enoate proposed by the ChemMLM

(c) False positive with the side product ammonia proposed by the ChemMLM and help specie benzene proposed by the ChemBalancer

Figure 7. Examples of false positive complete reactions: (a) a false positive ChemBalancer completed reaction, (b) a false positive ChemMLM completed reaction, and (c) a false positive reaction with both ChemBalancer and ChemMLM added molecules. The added molecules are shown in the red dashed boxes. Atom mapping cannot be identified for the reactions in (b) and (c).

The ChemMLM-completion results

ChemMLM learned from the limited complete reactions and broadened the boundary of complete reaction space to some originally incomplete reactions, still retaining the errors and noise present in the available complete reactions.

Model training results

The ChemMLM models were trained from two data sources, the USPTO reactions and the combined reactions with Reaxys®. Several attempts to tune training arguments were conducted for the ChemMLM learned from USPTO itself, whilst only one set of training arguments was conducted on the larger combined reaction dataset, with the model configuration and the most

https://doi.org/10.26434/chemrxiv-2023-hrgfw ORCID: https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0

optimal training arguments stated in Section “The ChemMLM model training and prediction”. As shown in Figure 8, using a large transformer model with approximately 83 million parameters to learn the large combined reaction dataset with 7 million reactions, the model took approximately 25 days to complete 17 epochs of training, with each colour in the training and validation loss curves corresponding to a single training day. The training and validation losses approach 0.121 and 0.118 eventually. As shown in the trend of training and validation loss curves, the losses have not converged yet, and the learning process could continue. However, it was interrupted since this long training process could not reach as good performance as the ChemMLM model trained from the smaller USPTO dataset in six days, which stopped at 0.050 and 0.052 training and validation losses respectively after 125 epochs. This was stopped because the validation loss starts to be higher than the training loss. Potentially the combined dataset contains a higher level of semantic information, which could possibly train a better ChemMLM model. However, the ChemMLM model learned from the larger combined dataset was not fine-tuned since this was computationally too expensive. The larger learning rate and batch size, as shown in Table 2, were chosen for the model trained from the combined dataset for the purpose of faster training.

https://doi.org/10.26434/chemrxiv-2023-hrgfw ORCID: https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0

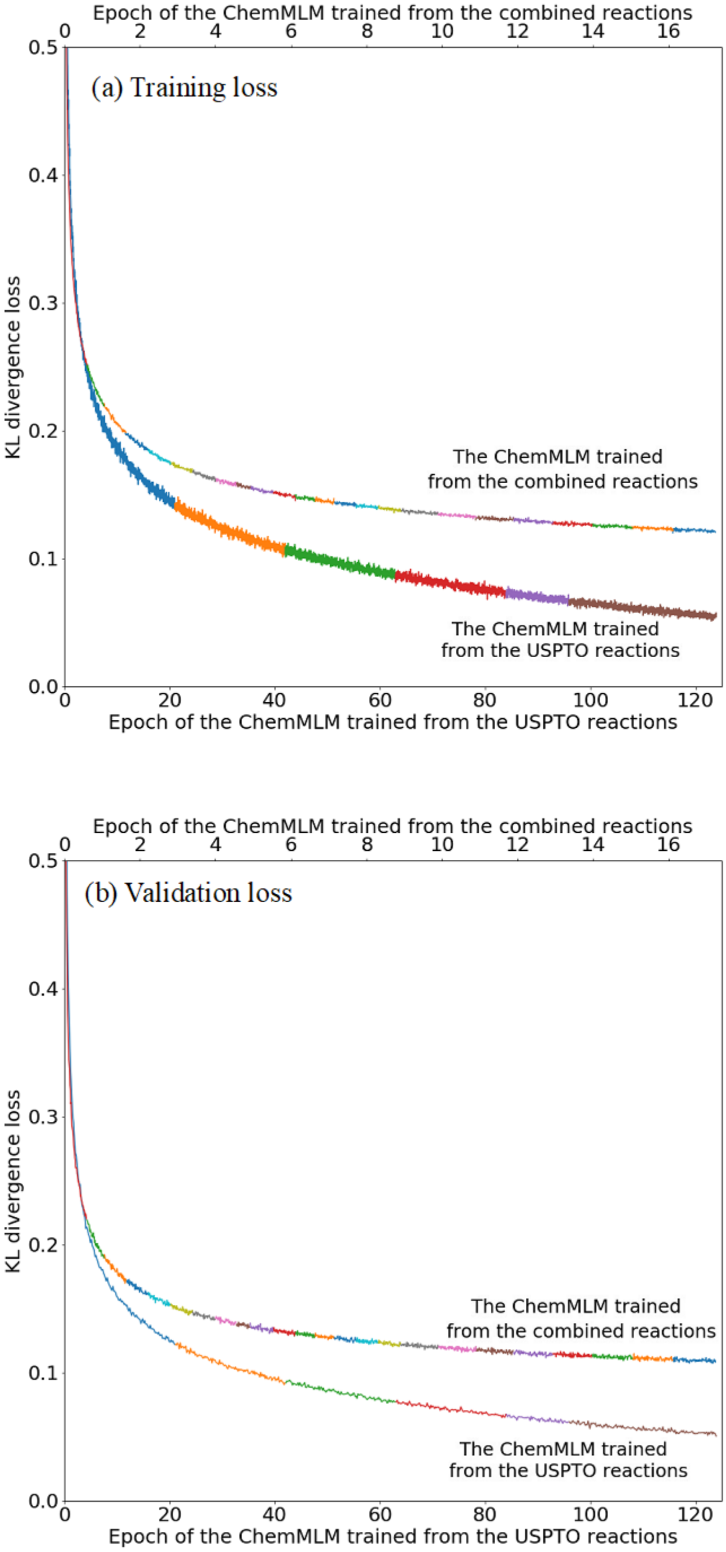


Figure 8. The training and validation KL divergence losses of the ChemMLM model learned from (a) USPTO reaction dataset only, and (b) combined reaction dataset. The figure only shows the most optimal model using the training argument set stated in Table 2.

Model test results

https://doi.org/10.26434/chemrxiv-2023-hrgfw **ORCID:** https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

The models were assessed from the test dataset of the two sources to determine the ratios of correct prediction of the exact hidden molecules, at both sides of reaction, as shown in Table 4.

Table 4. The ratios of correct prediction of the exact hidden molecules with short, middle and long token lengths at the LHS and the RHS of the reactions in the test reaction dataset, for the ChemMLM models trained from two complete reaction data sources, USPTO and combined reaction datasets.

Models	Molecule length	USPTO	Combined
LHS ChemMLM	Short	100%	100%
	Middle	75.2%	65.7%
	Long	24.5%	1.8%
RHS ChemMLM	Short	99.8%	100%
	Middle	81.3%	62.4%
	Long	8.2%	4.9%

From the test dataset, the model' ability to predict short token length (token length=1), middle token length (1<token length≤10) and long token length (token length>10) hidden molecules were assessed, and the comparisons between exemplified predicted and ground truth hidden molecules at different lengths are shown in Table 5. The ChemMLM models could predict almost 100% the ground truth hidden molecules at the short token length, while remain very high correct rate at middle token lengths. However, the models were unable to predict long token length molecules, with predictions not only semantically incorrect, but also invalid in SMILES syntax. For example, the last predicted molecule example shown in Table 5 has more right parentheses than left. This is because with longer token lengths in the hidden molecules, it becomes harder for the ChemMLM model to pick up neighbour semantic context of the reaction SMILES strings, since the very neighbour tokens are also within the hidden molecule and uncertain.

Table 5. A comparison between the exemplified ground truth hidden molecules and the molecules at different token lengths predicted by the ChemMLM models trained from USPTO complete reactions.

Molecule length	Prediction	Ground truth
Short	Cl	Cl

https://doi.org/10.26434/chemrxiv-2023-hrgfw ORCID: https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0

	O	O
	CO	CO
	CCCCO	CCCCO
Middle	[H][H]	[H][H]
	CC(=O)O	CC(=O)O
	NC(=0)c1ncc(Br)cc1N	NC(=0)c1ncc(Br)cc1N
	O=C(CCl)Nc1	O=C(CCl)Nc1ccnnc1
Long	CC=C(O)C1C)C(C(C)c1ccccc1	CC1=C(c2ccccc2)OC(C)(C)C1=O
	0=C1CC1c2)2)c)c2c3ccccc3c2ccccc1	O=c1cc(-c2ccccc2)c2ccc3ccccc3c2[nH]1
	CCCCCCc1cn(CCS(=O)(=O)cc(CCccccc2)nn1	CCCCCCc1cn(CCS(=O)(=O)c2ccc(C)cc2)nn1
	CNC(=O)/C=C/[C1Cc1HCCCCCCCC))))))))))(=))OCc1ccccc1	CNC(=O)/C=C/[C@H](Cc1ccccc1)NC(=O)[C@H](CCCNC(=O)OCc1ccccc1)NC(C)C

The longer token length molecules are longer than “a missing word” under the definition of the MLM. It fits better with another application of the BERT transformer model, the next sentence prediction,[26, 27] which predicts the entire next sentence based on the previous context. This could be an area of exploration for future prediction of longer molecules. However, this has not been implemented in this paper, since the low model predictability of longer molecules would not significantly affect the ChemMLM’s ability to complete reactions. The missing molecules in an incomplete reaction are commonly the side reactants and side products, which are usually smaller molecules.

Table 4 also shows the ChemMLM models do not show significant difference in prediction of hidden molecules between the LHS and RHS, in terms of their correction rate of short, middle and long token length molecules. This is because the models were trained using a dynamic mask method, which learned the semantics at two sides of the reactions simultaneously. However, the ChemMLM model trained from USPTO data shows better model performance in test data compared with the current ChemMLM model learned from the combined data. Perhaps a fine-tuned ChemMLM model learned from the larger reaction dataset could increase correct prediction rate in the test dataset, but the ChemMLM learned from the USPTO reactions were

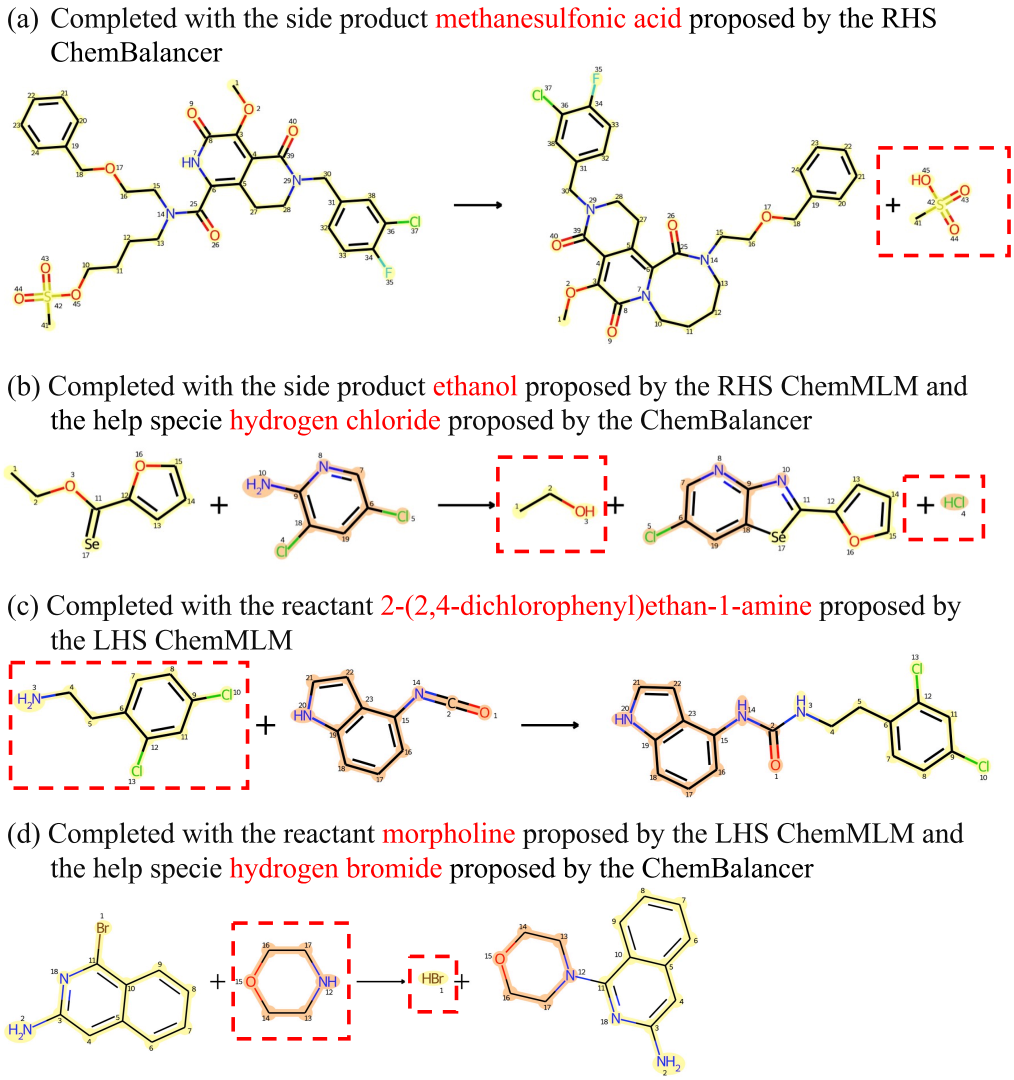
https://doi.org/10.26434/chemrxiv-2023-hrgfw ORCID: https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0

sufficiently predictive to predict the short and middle length hidden molecules. This indicates that the USPTO reaction dataset is capable to provide comprehensive types of complete reactions for the ChemMLM model to learn from. The ChemMLM model in the rest of this paper is referred to the ChemMLM trained from the USPTO reaction only.

Reaction completion results

For the incomplete reactions, ChemMLM proposes missing molecules either at the LHS or RHS based on ChemBalancer suggestions of LHS or RHS species insufficiency, and the proposed solutions are further checked by ChemBalancer. The exemplified reactions completed in this workflow are shown in Figure 9. In these examples, (a) and (c) show examples of reaction completion with the molecules proposed by ChemMLM itself, whilst (b) and (d) show examples of reaction completion by the molecules proposed by ChemMLM. All the exemplified reactions were manually evaluated to be correct completions based on reaction mechanisms.

https://doi.org/10.26434/chemrxiv-2023-hrgfw **ORCID:** https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0


(a) Completed with the side product methanesulfonic acid proposed by the RHS ChemBalancer

(b) Completed with the side product ethanol proposed by the RHS ChemMLM and the help specie hydrogen chloride proposed by the ChemBalancer

(c) Completed with the reactant 2-(2,4-dichlorophenyl)ethan-1-amine proposed by the LHS ChemMLM

(d) Completed with the reactant morpholine proposed by the LHS ChemMLM and the help specie hydrogen bromide proposed by the ChemBalancer

Figure 9. Examples of ChemMLM reaction completion results: (a) a RHS insufficient reaction completed by the RHS ChemMLM only, (b) a RHS insufficient reaction completed by the RHS ChemMLM plus ChemBalancer, (c) a LHS insufficient reaction completed by the LHS ChemMLM only, and (d) a LHS insufficient reaction completed by the LHS ChemMLM plus ChemBalancer. All reactions are atom mapped by RXNMapper[9] and labelled with atom mapping indices.

However, the reaction completion by ChemMLM still retains false positive reactions, with exemplified reactions shown in Figure 7(b) and (c). In Figure 7(b), the reactant is supposed to hydrolyse into two fragments, whilst the RHS ChemMLM predicts a chemically meaningless side product cyclohexyl 2-methylprop-2-enoate, and this updated reaction could be materially balanced with stoichiometric coefficients by ChemBalancer. Similarly, in Figure 7(c), an incorrect side product ammonia was proposed by the RHS ChemMLM. With the help species

https://doi.org/10.26434/chemrxiv-2023-hrgfw ORCID: https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0

benzene proposed by ChemBalancer, the updated reaction could be materially balanced with stoichiometric coefficients. These false positive complete reactions were caused only in very rare situations. This happens when some of the 333 enumerations to predict the missing molecules by ChemMLM were not semantically meaningful but their SMILES syntax was valid. This prediction remains valid only when the updated reactions were possible to be balanced by large values of stoichiometric coefficients in coincidence.

For reaction completion, all incomplete reactions in USPTO dataset from the ChemBalancer subunit were passed to ChemMLM for missing molecule prediction, whilst only 1% from Reaxys® was randomly sampled and fed in ChemMLM. This is because the missing molecule prediction by ChemMLM was computationally expensive, which makes it harder to process the larger scale Reaxys® reactions data. Statistics for the reaction completion results are shown in Table 6, and this was concluded from all USPTO reactions and sampled Reaxys® reactions. From Table 6, it is seen that a higher percentage of incomplete reactions could be completed from the USPTO than the Reaxys® reactions. This is very likely because a higher percentage of reactions in Reaxys® were completed by ChemBalancer, as shown in Table 3, leaving higher difficulty for ChemMLM to predict missing molecules. Also, in general, the LHS species insufficient reactions were harder to be completed by this workflow compared with the RHS species insufficient reactions.

Table 6. Statistics for the LHS and RHS species insufficient reactions completion results by ChemMLM for the reaction records in USPTO and Reaxys® databases.

Reaction database	USPTO	Reaxys®
Completed reactions from LHS insufficient reactions	24.1%	19.7%
– Completed by ChemMLM itself	8.2%	6.7%
– Completed by ChemMLM & ChemBalancer reagents	0.4%	4.9%

https://doi.org/10.26434/chemrxiv-2023-hrgfw ORCID: https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0

– Completed by ChemMLM & ChemBalancer help species	15.5%	8.1%
Completed reactions from RHS insufficient reactions	42.7%	33.4%
– Completed by ChemMLM itself	3.7%	11.9%
– Completed by ChemMLM & ChemBalancer reagents	0.5%	3.2%
– Completed by ChemMLM & ChemBalancer help species	38.5%	18.3%

Summarised from Table 3 and Table ***6***, 3.3% and 7.2% of the total reactions from USPTO and Reaxys® are originally complete/balanced reactions, respectively. 17.6% and 44.1% of total reactions in USPTO and Reaxys® were completed respectively by the first method, ChemBalancer. Afterwards, 34.8% and 16.8% of reactions were completed with the missing molecule prediction by the second method, ChemMLM. In total, 52.4% of USPTO and 60.9% of Reaxys® were completed by the entire proposed method, from which the Reaxys® percentage is an estimated number, since only 1% of LHS and RHS species insufficient reactions were sampled and fed to the ChemMLM model. Moreover, a small portion of the completed reactions are false positive completions, and those reactions could not be differentiated currently unless they were manually removed from the completed reactions.

Fragment method results

This section discusses a potential future method to solve reaction completion. Using an atom mapping tool to understand the atom rearrangements from the reactants to the products, the unmapped fragments in the reactants could be detected. These fragments themselves could be potential missing products of the incomplete reactions, coming from the cleavage of the reactants, or could be recombined to form new product molecules. Figure 10(a) shows an example reaction completed by this method. Using the atom mapping tool RXNMapper,[9] it is known that in the anion of the reactants, only the carbon atom, marked as "C[:1]", is mapped

https://doi.org/10.26434/chemrxiv-2023-hrgfw **ORCID:** https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

into the products. Therefore, it is very likely that the carbon-carbon bond connected to the carbon atom is broken to form another anion product shown in the red box in Figure 10(b). The complete reaction in Figure 10(b) has all atoms in reactants mapped into product by the RXNMapper.

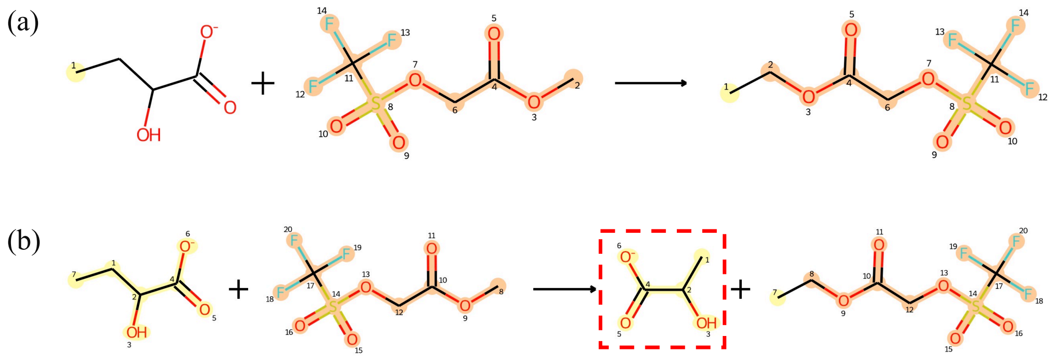


Figure 10. An example of a reaction completion by the proposed fragment method: (a) the original reaction recording, and (b) the complete reaction with the proposed product from unmapped fragment.

However, currently this fragment method cannot be applied to all reaction records but demonstrated here as a potential future method. This is because firstly, there is no perfect atom mapping tool which always gives ground truth mapping results, from which the unmapped fragments are always uncertain. Secondly, the unmapped fragments could undergo further cleavage or recombination steps, which at this stage could only be manually examined. With further development of this method in the future, this could be a third-tier method if the reaction completion fails from ChemBalancer and ChemMLM predictions.

Conclusions

Current reaction data records from multiple reaction databases including USPTO and Reaxys® are missing important reaction participating species and stoichiometric coefficients. Completing reaction structures in reaction data would make it possible to understand the true molecular flow for the reaction routes in CASP tasks. Although multiple tools such as rule-

https://doi.org/10.26434/chemrxiv-2023-hrgfw **ORCID:** https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

based reaction templates, atom mapping software, *etc* have been developed to investigate molecular transformations in reactions, identifying the missing molecules in reactions remains a challenging task, since most of these data-driven tools were developed from the incomplete reaction datasets. The reaction completion challenge could not be fully solved unless a "golden dataset" with comprehensive complete reactions appears. With the limited complete reaction data sources, the objective in this work was to learn from these limited complete reactions and broaden the boundary of the complete reaction space to some originally incomplete reactions, towards the reaction structure completion of the current reaction records.

To do this, a workflow including both heuristics and machine learning methods was proposed. From USPTO and Reaxys® reaction data, a heuristic-based balancing algorithm, namely, ChemBalancer was developed to investigate if a reaction is balanced based on atom equality of each element and electron balance at the two sides of the reaction. ChemBalancer also attempted to complete the imbalanced reactions by adding molecules from reagents, solvents and a list of most frequently appeared small molecules in reactions. The remaining incomplete reactions were classified as either LHS or RHS species insufficient reactions. ChemBalancer identified 3.3% and 7.2% originally complete/balanced reactions from USPTO and Reaxys® respectively, and further completed 14.3% and 36.9% of the total reactions by adding possible missing molecules to the reactions.

A machine learning BERT transformer model, namely, ChemMLM was developed to learn from the semantic meaning of the reaction SMILES strings, from the complete reaction dataset identified from the last step. Using the masked language model scheme, analogous to missing words in sentences, ChemMLM was trained to predict the possible missing molecules for the

https://doi.org/10.26434/chemrxiv-2023-hrgfw **ORCID:** https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

incomplete reactions. From the test dataset, it is proven that ChemMLM was confident to predict short and middle (token length<=10) length missing molecules. Given the classification of LHS or RHS species insufficient of the incomplete reactions, ChemMLM either predicted missing molecules to the LHS or RHS. ChemMLM further completed 34.8% and 16.8% reactions from the total USPTO and Reaxys® reaction dataset. The entire workflow completed 52.4% and 60.9% of reactions from USPTO and Reaxys® in total, retaining with false positive complete reactions that could be only manually detected.

For future works, to improve the prediction accuracy of longer length molecules, as a longer length molecule (token length>10) can have molecular SMILES strings analogous to a sentence, the next-sentence prediction scheme of the BERT transformer could potentially be used to predict these molecules. However, since most incomplete reactions are missing side reactants and products which are smaller molecules, this would not significantly increase the reaction completion rate. Moreover, to minimise the false positive rate in the completed reactions, a calibration step could be added. Atom mapping confidence score would be an important index to understand the atom rearrangement in reactions. However, since there is not a ground truth atom mapping method, this index could only be used as recommendation for false positive detection. A more promising atom mapping tool is required to calibrate the model completed reactions. Eventually, to ultimately achieving reaction structure completion, a “golden dataset” is called for. This requires a more accurate NLP text-mining tool to reduce the noise when mining reactions from literature, and manual curation of at least a few reactions from each reaction subcategory.

https://doi.org/10.26434/chemrxiv-2023-hrgfw **ORCID:** https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

Acknowledgements

CZ is grateful to Cambridge Trust CSC Scholarship for funding their PhD study. AA is grateful to Cambridge Centre for Advanced Research and Education in Singapore (CARES Ltd) for funding his PhD scholarship via the NRF CREATE programme Cambridge Centre for Carbon Reduction Technologies (C4T), and to Chemical Data Intelligence (CDI) Pte Ltd for co-funding this studentship. We gratefully acknowledge collaboration with RELX Intellectual Properties SA and their technical support, which enabled us to mine Reaxys®.Copyright © 2020 Elsevier Limited except certain content provided by third parties. Reaxys® is a trademark of Elsevier Limited. Reaxys® data were made accessible to our research project via the Elsevier R&D Collaboration Network.

Data Availability

Reaxys® molecule and reaction data are accessible to users via Elsevier. The code used for the ChemBalancer and the ChemMLM can be found online at: https://github.com/chonghuanzhang/balancing_rxn.

References

1. Elsevier Reaxys. https://www.reaxys.com/ (accessed 6 Feb).
2. Lowe, D. Chemical reactions from US patents. https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873 (accessed 2 May).
3. Weber, J. M.; Guo, Z.; Zhang, C.; Schweidtmann, A. M.; Lapkin, A. A., Chemical data intelligence for sustainable chemistry. *Chem. Soc. Rev.* **2021,** *50* (21), 12013-12036.
4. Kearnes, S. M.; Maser, M. R.; Wleklinski, M.; Kast, A.; Doyle, A. G.; Dreher, S. D.; Hawkins, J. M.; Jensen, K. F.; Coley, C. W., The Open Reaction Database. *J. Am. Chem. Soc.* **2021,** *143* (45), 18820-18826.
5. Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A., Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem.* **2016,** *55* (20), 5904-5937.
6. Synthia Synthia Organic Retrosynthesis Software. https://www.sigmaaldrich.com/GB/en/technical-documents/technical-article/chemistry-and-synthesis/organic-reaction-toolbox/resources#ref2 (accessed 09 Feb).
7. Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F., Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017,** *3* (5), 434-443.

https://doi.org/10.26434/chemrxiv-2023-hrgfw **ORCID:** https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

8. 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), 1434.
9. Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T., Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances 7* (15), eabe4166.
10. Nugmanov, R. I.; Mukhametgaleev, R. N.; Akhmetshin, T.; Gimadiev, T. R.; Afonina, V. A.; Madzhidov, T. I.; Varnek, A., CGRtools: Python Library for Molecule, Reaction, and Condensed Graph of Reaction Processing. *J. Chem. Inf. Model.* **2019,** *59* (6), 2516-2521.
11. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics.: Minneapolis, Minnesota, 2019; Vol. 1, pp 4171–4186.
12. Arun, A.; Guo, Z.; Sung, S.; Lapkin, A. A., Reaction Impurity Prediction using a Data Mining Approach**. *Chemistry–Methods* **2023,** *n/a* (n/a), e202200062.
13. Dahlgren, B. r., ChemPy: A package useful for chemistry written in Python. *The Journal of Open Source Software* **2018,** *3* (24), 565.
14. Salazar, J.; Liang, D.; Nguyen, T. Q.; Kirchhoff, K., Masked language model scoring. *arXiv preprint arXiv:1910.14659* **2019**.
15. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V., Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* **2019**.
16. 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 Cent. Sci.* **2019,** *5* (9), 1572-1583.
17. Gage, P., A new algorithm for data compression. *C Users J.* **1994,** *12* (2), 23–38.
18. Sennrich, R.; Haddow, B.; Birch, A. In *Neural Machine Translation of Rare Words with Subword Units*, Berlin, Germany, August; Association for Computational Linguistics: Berlin, Germany, 2016; pp 1715-1725.
19. 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. *Chem. Sci.* **2020,** *11* (1), 154-168.
20. Face, H., *The AI community building the future*. 2022.
21. Google, *Google Colab*. 2022.
22. Hendrycks, D.; Gimpel, K., Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* **2016**.
23. RDKit: Open-source cheminformatics. http://www.rdkit.org (accessed 6 Feb).
24. Lowe, D. M. Extraction of chemical structures and reactions from the literature. University of Cambridge, 2012.
25. Pinner Reaction. In *Comprehensive Organic Name Reactions and Reagents*, John Wiley & Sons, Ltd: 2010; pp 2237-2240.
26. Sun, Y.; Zheng, Y.; Hao, C.; Qiu, H., NSP-BERT: A Prompt-based Zero-Shot Learner Through an Original Pre-training Task--Next Sentence Prediction. *arXiv preprint arXiv:2109.03564* **2021**.
27. Shi, W.; Demberg, V. In *Next sentence prediction helps implicit discourse relation classification within and across domains*, Proceedings of the 2019 conference on empirical

https://doi.org/10.26434/chemrxiv-2023-hrgfw **ORCID:** https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), 2019; pp 5790-5796.

https://doi.org/10.26434/chemrxiv-2023-hrgfw **ORCID:** https://orcid.org/0000-0001-7621-0889 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0