New Application of Natural Language Processing（NLP）for Chemist: Predicting Intermediate and Providing an Effective Direction for Mechanism Inference

Jiangcheng Xu[a,b]+, Yun Zhang[a]+, Jiale Han[a]+, Haoran Qiao[c], Chengyun Zhang[a], Jing Tang[a], Xi Shen[a], Bin Sun[a], Silong Zhai[a], Xinqiao Wang[a], Yejian Wu[a], Weike Su*[a], Hongliang Duan*[a]

These authors contributed equally: Jiangcheng Xu, Yun Zhang, Jiale Han

[a] J. Xu, Y. Zhang, J. Han, C. Zhang, J. Tang, X. Shen, B. Sun, S. Zhai, X. Wang, Y. Wu, Prof. W. Su*, Prof. H. Duan*
Artificial Intelligence Aided Drug Discovery Institute, College of Pharmaceutical Sciences;
Collaborative Innovation Center of Yangtze River Delta Region Green Pharmaceuticals
Zhejiang University of Technology
Hangzhou, 310014 (P. R. China)
E-mail: Pharmlab@zjut.edu.cn; hduan@zjut.edu.cn

[b] J. Xu
Hangzhou Vocational & Technical College
Hangzhou, 310014 (P. R. China)

[c] H. Qiao
College of Mathematics and Physics
Shanghai University of Electric Power
Shanghai, 201203 (P. R. China)

Supporting information of this article can be found under:

Abstract: Predicting and proposing the reaction mechanism, as well as speculating the reaction intermediates are great challenges among the development of modern organic chemistry. Herein, a model from Natural Language Processing (NLP) was firstly employed to learn and perform the task of intermediate prediction, which is served as a language translation task. Radical cascade cyclization is prevalently used in life science and pharmaceutical projects, while the regioselectivity of radical attack is difficult to predict. The model is trained on self-built dataset to tackle the challenge. And transfer learning was used to surmount the restriction of limited amounts of data. The NLP transformer model performs well with remarkable accuracy, providing an efficient instruction for mechanism understanding. Manual encoding of rules is not required, thus, providing a favorable tool towards solving the challenging problem of computational organic chemical mechanism inference.

Introduction

The molecules and materials design significantly depend on the understanding the exact mechanism of a reaction, such as the order of bonds formation and cleavage. For decades, chemists have picked apart chemical reactions to have an in-depth understanding of each steps involved with the aim of discovering new chemistry[1], and optimization of chemical reactions[2]. However, determination of a reaction mechanism usually needs integrate a variety of information from indirect observation, whose intermediates are hardly to be observed during the transformation.

Significant efforts have been made on the journey to explore the mechanism of chemical reactions.[3-4] The inference of mechanism using manually composed transformation rules have been developed along the years, and can be applied in reaction prediction.[5] However, these methods require manual encoding, as long as new reactions are discovered, old projects have been already outdated. Other methods are based on the physical calculations, such as density functional theory (DFT).[3,6] However, these approaches usually require high computational cost and rely on experienced chemists.

Recently, chemist has witnessed several successful applications of artificial intelligence (AI) algorithms.[7-17] As an important area of AI, Natural Language Processing (NLP)[18] models emerged as robust and effective approaches in the field of organic chemistry, showing promising results in reaction prediction[19-22], retrosynthesis[23-27] planning. However, in the field of chemical reactions mechanism inference, relevant studies are rarely reported.

The mechanism inference of organic chemistry is more like an art: the intermediate formed in each transformation varies a lot, and is difficult to be characterized. Thus, in face of a complex organic reaction, it seems that there are a thousand “Hamlets” in a thousand chemists’ eyes. The domain complexity and lack of sufficiently curated data hindered further technological developments of NLP model in this field. Encouraged by the breakthrough of artificial intelligence, the feasibility of application of NLP model in speculating mechanism of organic chemistry was discussed in this work.

Herein, radical cascade cyclization was selected as the object to explore the reaction mechanism by NLP transformer model. Recently, radical chemistry has become a heavily investigated research field, and radical cascade cyclization is highly valuable for preparation of cyclic compounds since the processes are performed in a single preparation step to construct carbo- and heterocycles,[28-30] which often show interesting biological properties with potential use in medicinal chemistry (Fig 1).[31-33] In this work, there are two additional reasons for choosing the topic of radical cascade cyclization. Firstly, the regioselectivity of radical attack in radical cascade cyclization is difficult to predict even for experienced chemists, therefore, a convenient method to explore the reaction process is urgently needed. Secondly, comparing with the double-electron reaction, the intermediate of radical reaction can be captured and verified. Accordingly, the correct data for training is more likely to be gathered from literatures.

The quality and scale of data play a key role in the model. Information about chemicals can be found in databases such as PubChem,[34] ChEMBL,[35] Reaxys,[36] SciFinder[37] databases and so on.[38] Unfortunately, to the best of our knowledge, information about intermediate is not recorded in above databases. Therefore,

we gathered data from related literature and constructed a data set with intermediate involved. Besides the shortage of original experimental data, chemical data collection is sophisticated and time-consuming, leading to the dilemma of small data volume. To our delight, our research group has carried out some prediction work based on small data set via transfer-learning strategy, which proved a better accuracy of prediction.[39-42] We believe that the approach would expand the application for NLP model with limited data sets in the field of chemistry.

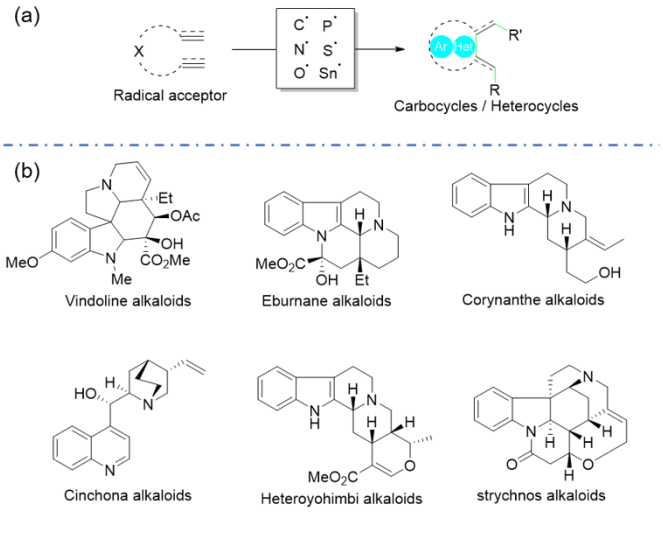

Figure 1. (a) Radical cascade cyclization is highly valuable for preparation of cyclic compounds since more than two bonds can be formed by a single preparation step. (b) Pharmaceuticals and complex natural products can synthesis via radical cascade cyclization.

In this study, NLP transformer model was firstly applied to learn and perform the intermediate prediction subtask (Fig 2). The general chemical reaction data set was used to pre-train the transformer model to learn general chemical knowledge. Afterwards, the transformer model is trained end-to-end on a self-built data set of radical cascade cyclization with corresponding intermediate from literature. Finally, for a given reaction equation, the model output the most reasonable intermediate. The result indicates that NLP transformer model performs well to predict intermediate, with a total accuracy up to 93.5%, and provides an effective direction for mechanism inference.

Results and Discussion

To begin with, key intermediate was defined for a better understanding of the reaction mechanism. With the help of the definition, radical cascade cyclization data set was established. Finally, NLP transformer model was trained for predicting key intermediate.

Radical cascade cyclization generally involves four key stages[29] (Fig. 3a). (i) radical formation: the process of single-electron transfer (SET). (ii) radical addition: the formation of a radical intermediate with radical attack on an unsaturated bond. (iii) radical cyclization: the formation of carbon-carbon/hetero-atom bond. (iv) radical quenching: radical intermediate is quenched by another radical donor or hydrogen abstraction. In this work, the intermediate generated after the first radical addition was defined as “Key intermediate Ⅰ”, which can show the regioselectivity of the first radical addition. Subsequently, the intermediate which was formed after the cyclization was defined as “Key intermediate Ⅱ” if the cascade reaction constructed only single carbo-/heterocycle. In the case of multiple rings constructed, the intermediate before the last cyclization step was defined as “Key intermediate Ⅱ”. Key intermediate Ⅱ can reflect important information whether rearrangement, hydrogen or aryl migration and other transformation occurred during the reaction (further information in section 1 of the ESI).

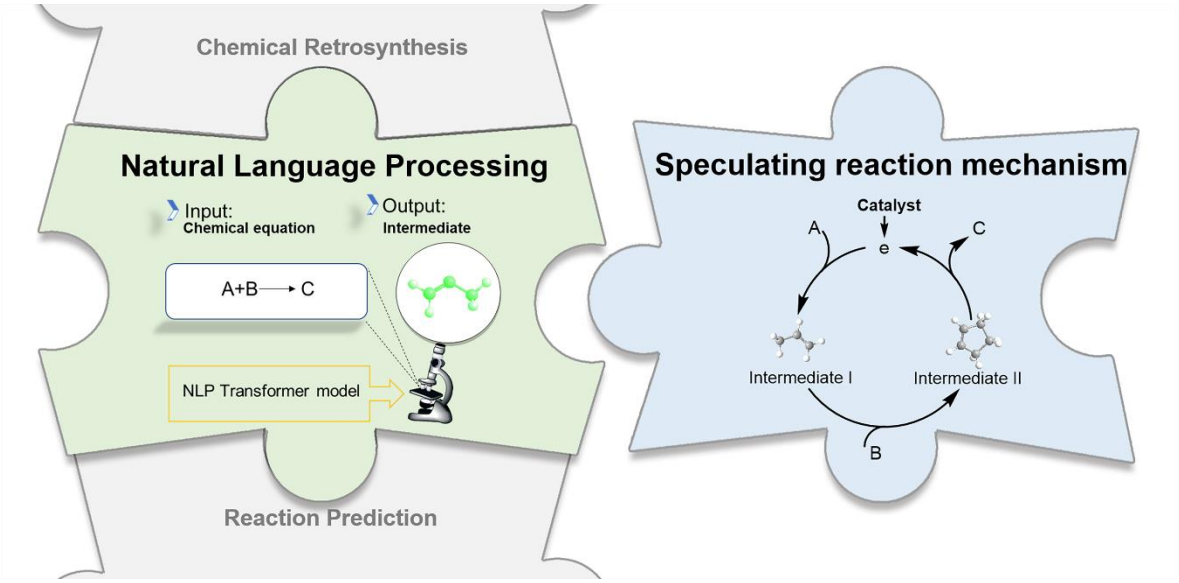

Figure 2. A proof-of-concept methodology for predicting intermediates of radical cascade cyclization. For a given reaction equation, the model simulated observation device in an indirect way that can assist chemists to speculate organic chemical reaction mechanism. In reaction prediction and retrosynthesis, artificial intelligence has done an excellent job on end-to-end prediction (from reactant to product and its inverse process). However, the same importance should be attached to figuring out what happened during a reaction for understanding the entire reaction.

In this work, speculating mechanism of radical cascade cyclization can be regarded as a maze game. In the maze game, if there is a “road sign”, players can quickly walk through the maze (Fig. 3b). Moreover, if there are multiple possible routes, the exact “road sign” can indicate the most reasonable route. The “key intermediate” serves as the “road sign” in the maze game of mechanism inference.

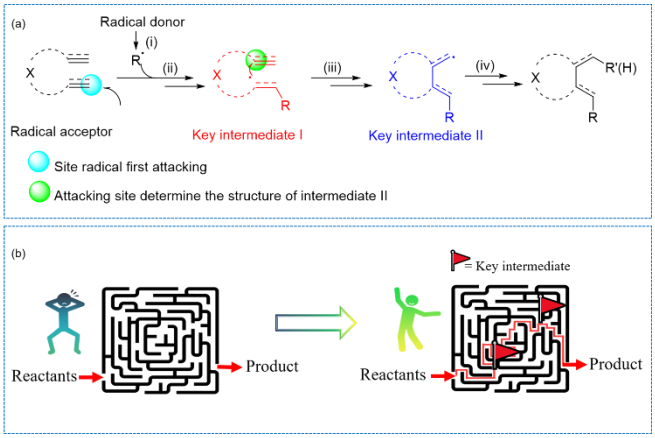

Figure 3. (a) The definition of key intermediate Ⅰ & Ⅱ of radical cascade cyclization in the background of four key stages. (b) Key intermediate Ⅰ & Ⅱ can effectively navigate organic chemical reaction mechanism inference.

Hundreds of literatures were collected and careful attention was paid to the latest achievement in the development of radical cascade cyclization methodologies, especially the synthesis of carbo- and heterocycles in this rapidly growing research field. Subsequently, we analyzed the mechanism inference in these literatures, especially focused on the mechanism verified by calculations (such as DFT), or based on experimental verification (kinetic isotope effect (KIE) or radical capture). In the following work, simplified molecular-input line-entry system (SMILES) was used to represent reaction equation collected with intermediate involved. Finally, all data were standardized utilizing RDkit (version 2019.03). The stereo-configuration information of reactant, product and intermediate were all retained. By this strategy, the data set of radical cascade cyclization is established manually (further information in section 2 of the ESI).

The self-built data set contains a diverse of information, including 874 radical cascade cyclization chemical equations and corresponding 1,748 key intermediates. Specifically, the number of newly constructed rings by radical cascade cyclization varies from single to multiple: 428 reactions construct single ring, 321 reactions construct 2 rings, and 125 reactions construct more than 3 rings (Fig 4a). Furthermore, the radical center varies in carbon-centered radicals (356 reactions), sulfur-centered radicals (266 reactions), phosphorus-centered radicals (104 reactions), nitrogen-centered radicals (92 reactions), and Tin-centered radicals (56 reactions) (Fig. 4b). At last, the formation of radical-

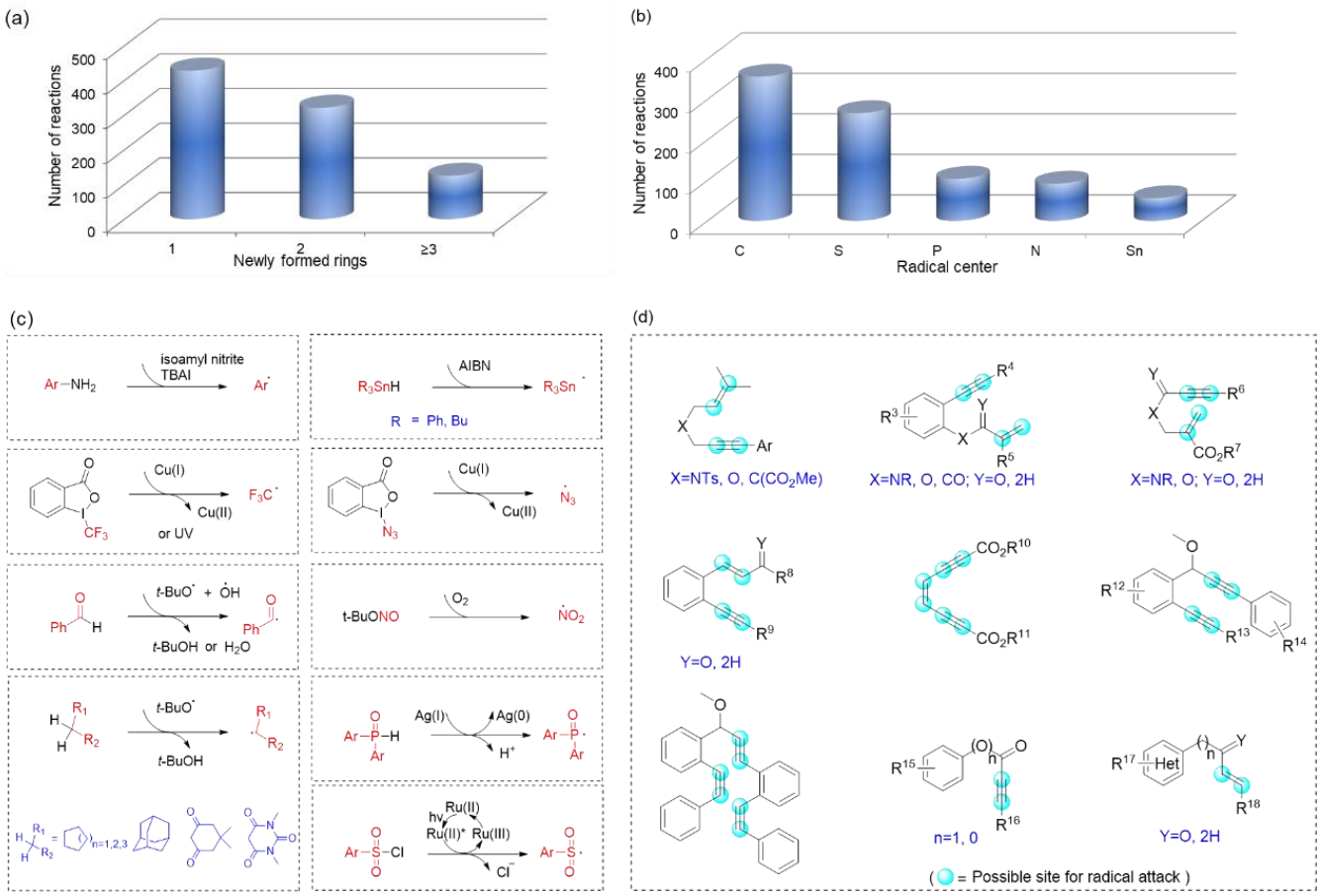

Figure 4 Diverse reactions in self-built data set. Distribution of (a) count of newly formed rings (b) types of radical center. (c) The most common way to generate radical center. (d) The most common radical acceptor with its possible site for radical attack.

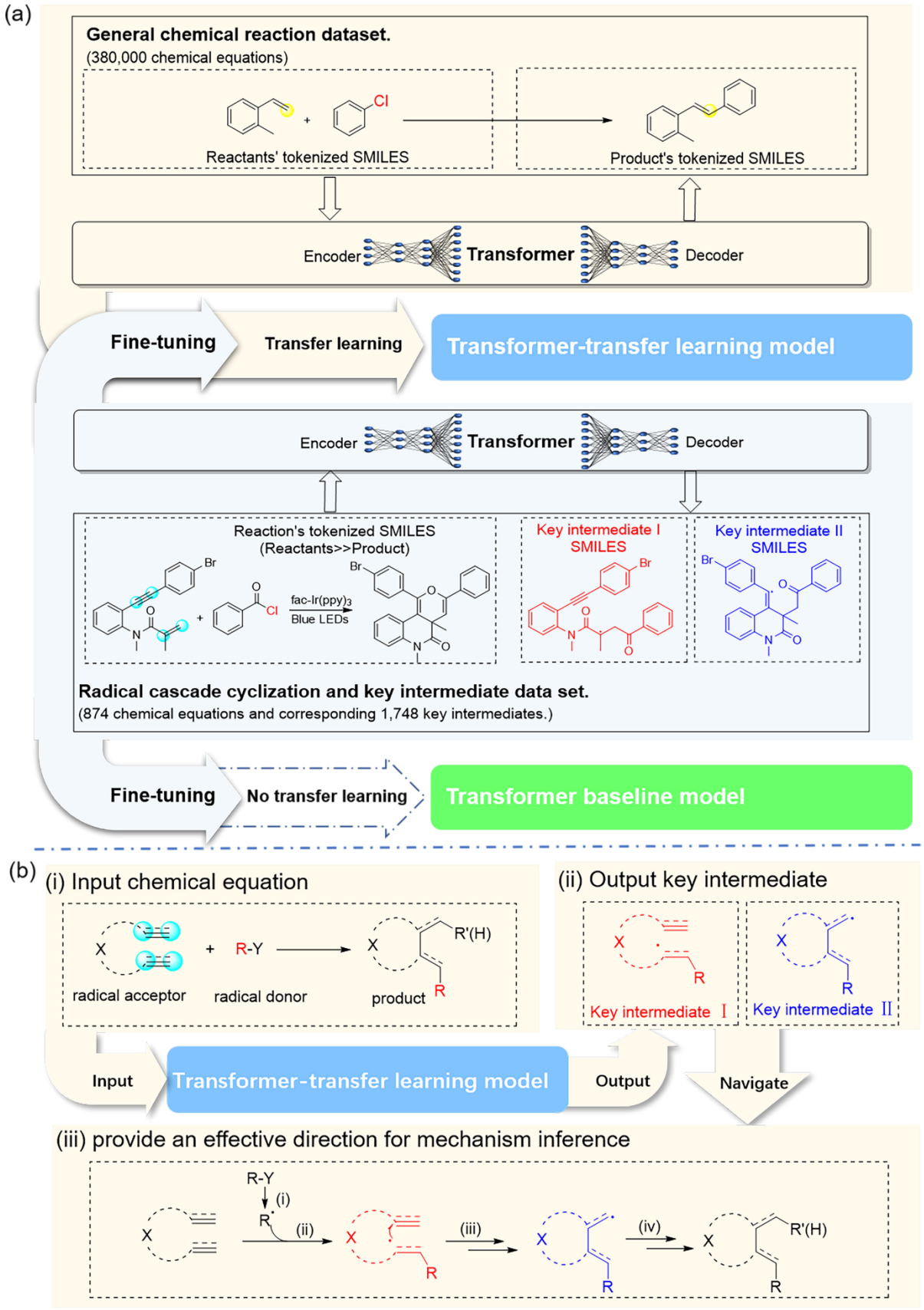

Figure 5. Schematic of the approach to predict radical cascade cyclization intermediate. (a) Comparison of transformer-transfer learning model and transformer baseline model. The transformer-transfer model is a plus version of transformer baseline model with knowledge learned from pre-training. Note that data features in pre-trained data set and in self-built data set are distinct. (b) The workflow of transformer model. With chemical equation inputted, transformer model predicts two key intermediates which would help chemist propose the most reasonable mechanism.

center varies because a donor can provide radical by different ways of SET (Fig 4c). As for radical acceptors, 1,n-diynes, 1,n-enynes, alkynyl (hetero) arenes are commonly used (Fig 4d).

The data set for pretraining, named general chemical reaction data set, contains approximately 380,000 chemical reactions. These reaction examples were originally sourced from Lowe's data set[43], which was extracted from United States Patent and Trademark Office (USPTO) patents, and then the reagents, solvent, temperature and other reaction conditions were deleted. Filtering work was done to clean duplicate, incorrect and incomplete reactions. Note that every single record in the general chemical reaction data set represents a reaction without intermediate, which is different from the self-built data set (Radical cascade cyclization and key intermediate). If we consider the intermediate as a special "product" containing radical, the forms of two data sets exist a huge similarity.

We try to treat intermediate prediction as a machine translation problem between simplified molecular-input line-entry system (SMILES) strings (a text-based representation) of chemical reaction equation and the key intermediates. Transformer model, a recent addition to Natural language processing (NLP) family, is the state-of-the-art model for reaction prediction and retrosynthesis.[20, 27, 44-45] Self-attention mechanism, feed forward network as well as multi-head attention makes transformer model an excellent tool for NLP tasks. Therefore, transformer model was applied to accomplish the task of predicting reaction intermediates (further information in section 3 of ESI).

Transfer learning, is generally exploited to adapt well-established source knowledge for learning tasks in weakly labeled or unlabeled target domain.[43-44] In this study, the process of pretraining on the general chemicals fulfill the initialization or feature extraction task in traditional machine learning. The basic chemical information and characteristics are applied to complete the target mission of predicting the key intermediate of radical cascade cyclization by the pretrained process (Fig 5a). After that, the self-built data set was randomly split into training, validation and test data set at a ratio of 8:1:1. To verify the improvement of fine-tuning, the performance was compared between transformer-baseline model without transfer learning and the one with transfer learning.

To measure the performance, we used accuracy over the predicting task, which was estimated with 10-fold cross-validation. Fig 6 and Table 1 show the accuracies of transformer-baseline and transformer-transfer learning models for radical cascade cyclization in different views. Transformer-transfer model exhibits the accuracy of 94.5% for key intermediate Ⅰ and 92.5% for key intermediate Ⅱ, which is much greater than the 29.7% and 26.6% for the transformer-baseline model (details can be found in Table S1-3 of ESI). Significantly, transformer-transfer learning model exhibited a much better performance. With the aid of the general chemical reactivity rules and knowledge obtained in the pretraining process, the transformer-transfer learning model was more accurate and be well used to cope with radical cascade cyclization's intermediate prediction. The result is consistent with our vision that the intermediate of radical cascade cyclization is special "product" of the reaction, while logic and knowledge of general chemical reactions are interlinked.

To figure out factors that affect the accuracy of intermediate prediction, we set a further exploration on different types of intermediate. For the convenience of analysis and understanding,

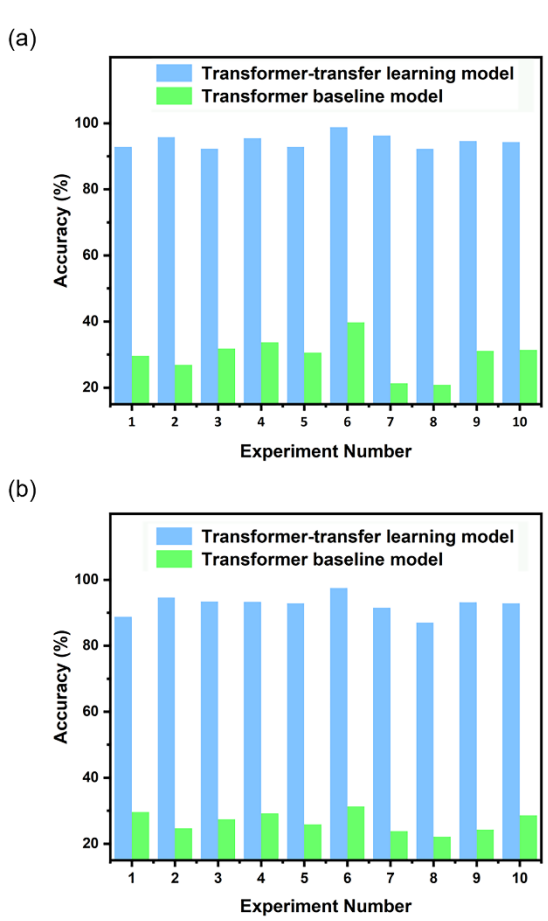

Figure 6. Comparison of the transformer-baseline and transformer-transfer learning model with accuracy of predicting key intermediate Ⅰ (a) and key intermediate Ⅱ (b). The result of 10 experiments indicated that transformer-transfer learning model has a significant advantage.

Table 1. The average accuracies of the transformer-baseline and transformer-transfer learning models for predicting key intermediates.

Model	Accuracy (%)	
	Key intermediate Ⅰ	Key intermediate Ⅱ
Transformer baseline model	29.7	26.6
Transformer transfer learning model	94.5	92.5

reactions were classified according to the number of newly constructed rings. Experiment 1 (refers to entry 1 in table S1 of the ESI) was selected as a representative example for analysis (99 reactions). The detailed accuracies of the transformer-transfer learning of different number of newly constructed rings are described in Table 2. In general, with the increasing of the constructed rings, the accuracy decreases. For key intermediate Ⅰ, as the number of rings increases, the prediction accuracy rates could be obtained with 94.2%, 92.5% and 83.3%, respectively. In some degree, the accuracy depends on the complexity of regio- or stereo-selectivity, and the training data decrease as the number of newly construct rings increases.

Table 2. The detailed accuracies of the transformer-transfer learning of different number of newly constructed rings.

Type of reaction[a]	Accuracy (%)		
	Key intermediate Ⅰ	Key intermediate Ⅱ	Key intermediate Ⅰ & Ⅱ [b]
1	94.2	93.8	90.4
2	92.5	82.5	77.5
≥3	83.3	66.7	66.7
Total	92.9	88.8	81.6

[a] Based on the number of newly constructed rings. [b] Predict two key intermediates simultaneously.

A similar result was obtained with respect to the intermediate Ⅱ, and the accuracy was 93.8%, 82.5%, and 66.7% as the number of rings increase. The accuracy of key intermediate Ⅰ is slightly higher than that of intermediate Ⅱ with accuracy of 92.9% and 88.8%, respectively. The possible reason is that the key intermediate Ⅱ is more complicated than the intermediate Ⅰ, increasing the difficulty of recognizing the reaction site. Moreover, we also trained the transformer model to predict two key intermediates simultaneously. Although the difficulty increased, the results showed no significant decrease on the performance with the prediction accuracy rate up to 81.6%. Expectedly, the accuracy decreased as the number of ring increase, furnishing the accuracy with 90.4%, 77.5%, and 66.7%, respectively. This showed the same orderliness as predicting key intermediate Ⅰ or key intermediate Ⅱ separately.

According to the Curtin-Hammett principle[48], the order of priority of attacking unsaturated bonds mainly depends on the structure of the radical acceptor. For the 1,n-enynes with non-terminal double bonds, the radical regioselectively attacked the α position of the alkyne, which may be due to the self-sorted kinetic process via the subsequent favored exo-trig radical cyclization (Fig 7a). Meanwhile, for this radical acceptor with the terminal alkenes, radicals underwent the selective addition to the activated alkene moiety of the enyne acceptor component (δ position) (Fig 7b), despite of a few special examples.[49] The regioselectivity is proposed by the fact that alkyne π-bond is stronger and less reactive than the π-bond of alkene, leading to the lower kinetic barrier of radical addition to the double bond.[50-51] Though it is difficult to summarize the rules for complex reaction, it can be learned by deep-learning model with suitable dataset in an indirect way.

Interpretability is essential for users to effectively understand, trust, and manage powerful artificial intelligence applications.[52-55] Specific functional groups, have an impact on the outcome of a reaction, even if they are far from the reaction center in the molecular graph (three-dimension) and therefore also in the SMILES string (one-dimension). It is worth mentioning that the transformer-transfer learning model exhibits a similar "thinking mode" to chemists'. A representative instance of correct predicting intermediates with the attention weights were shown in Fig 8a. Attention is the key to take into account complex long-range dependencies between multiple tokens.[56] On the way to predict key intermediate Ⅰ (Fig 8b), the transformer-transfer model initially focused on the structure of both reactants and product. Then the radical location focused on the unsaturated bonds of the radical attack and the nearby functional group structure and location, especially focus on the position of double bonds, since terminal alkene and non-terminal alkene will result in different intermediate. In this example, the token "!" (represent "start" or "end") prior to the "C=C" express the existence of terminal alkene, leads to radical attack at the terminal alkene.

The visualization of attention show radical location ("[ ]" represent radical) of the intermediate Ⅰ connect to this token coincidentally, which is similar to chemist's mental journey. While in the mission of predicting key intermediate Ⅱ, almost the same rule can be found that the structure of intermediate based on both reaction and product, and the function group nearby the alkyl (Fig 8c). More reaction predictions together with the attention weights detail, are found in the ESI.

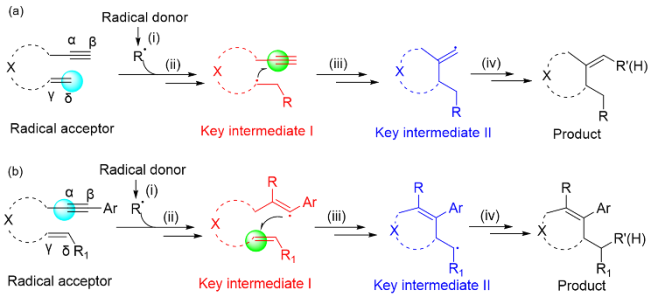

Figure 7. Plausible mechanism of radical cascade cyclization to 1,n-enynes. (a) For the radical acceptors 1,n-enynes with non-terminal double bonds, selective radical attacking at the α position of the alkyne. (b) For the terminal alkenes, regioselective addition of radical species at the δ position occurs firstly.

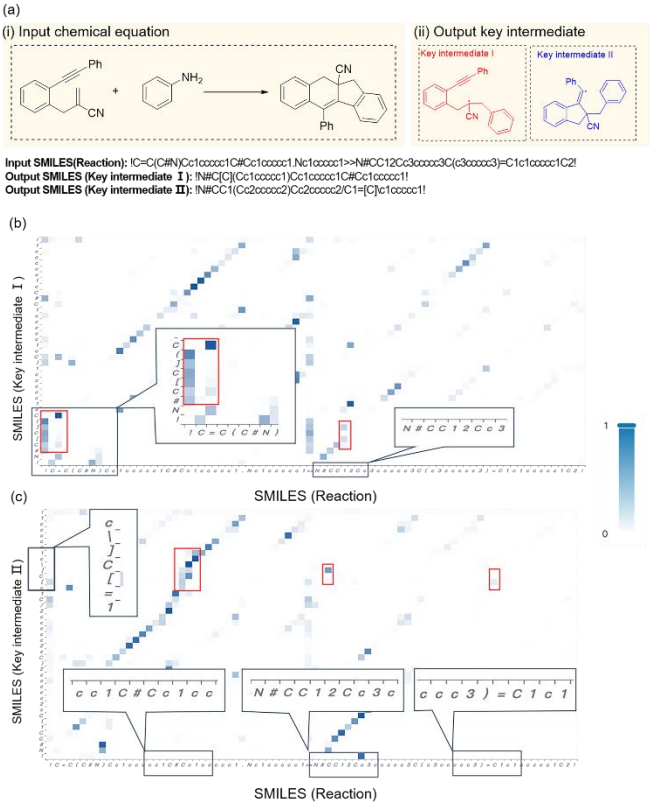

Figure 8 Attention weight interpretation. (a) Example randomly chosen from reactions in self-built data set. The visualization of attention weights for (b) key Intermediate Ⅰ and (c) key Intermediate Ⅱ.

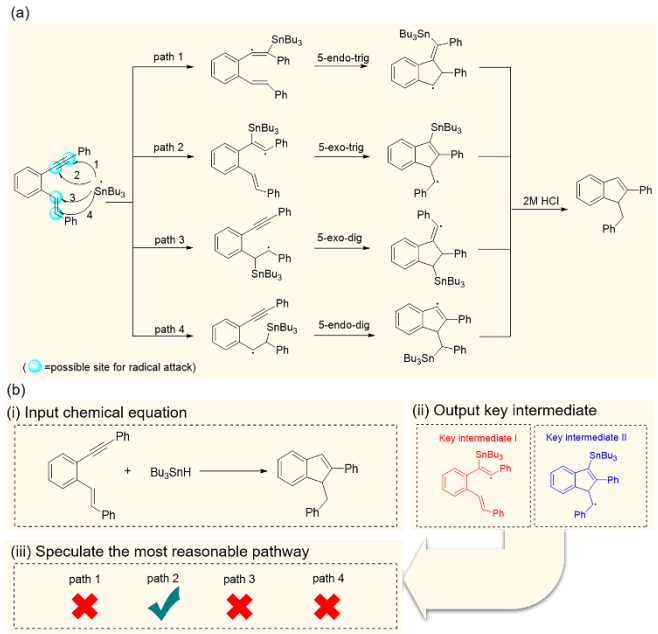

Figure 9. A representative example. (a) Regioselective attack at a disubstituted alkyne in the presence of electronically and sterically similar alkene is a great challenge. (b) With the aid of NLP transformer-transfer learning model, path 2 is inferred as the most reasonable pathway. The results exhibit a consistent result with computational analysis.

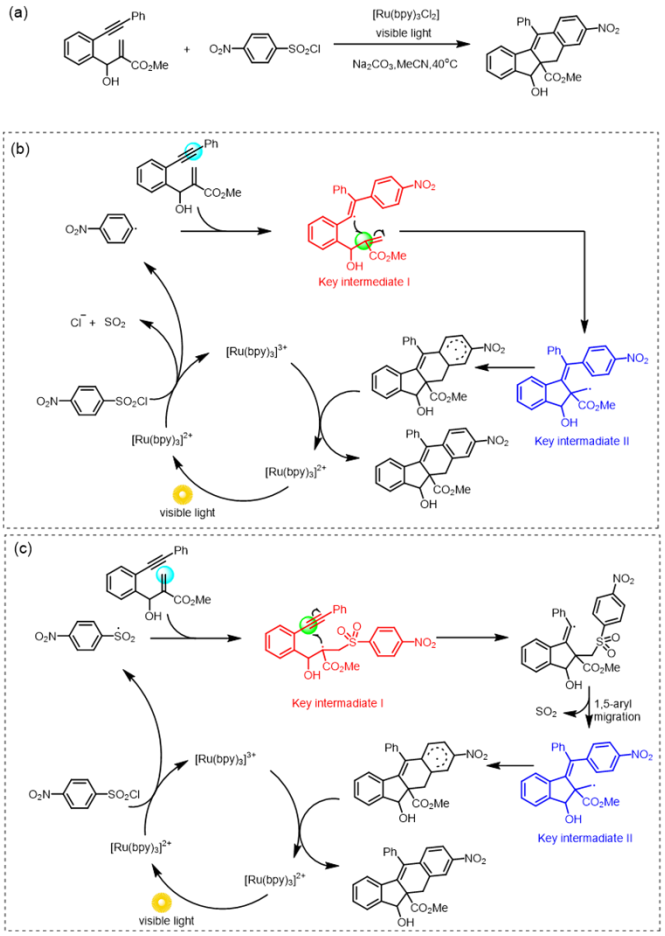

Figure 10. The NLP transformer-transfer learning model can revise some mechanism inference work in the reported literature. (a) Radical cascade cyclization of alkynes with alkenes and $ArSO_2Cl$. (b) Possible mechanism inferred by Li *et al*. (c) Another possible mechanism proposed by NLP transformer-transfer learning model, which is proved to be more reasonable by the further literature exploration.

Intermediate predicted can provide direction for mechanism inference. In the self-built dataset, enynes is the most common radical acceptor. There are two unsaturated bonds in enynes that participate in the radical cascade cyclization, which have at least four possible sites for radical attack. The Bu_3Sn-mediated radical cascade cyclization of aromatic enynes was taken as a representative example (Fig 9a), where the regioselectivity of radical addition is a great challenge in the presence of electronically and sterically similar alkene. Even for such complex prediction, the transformer-transfer learning model still achieved a good performance. After inputting chemical equation, the deep-learning model output two key intermediates, and suggested that the path 2 is the most reasonable route of this reaction (Fig 9b). Meanwhile, Alabugin *et al*.[57] reported a reasonable reaction path with the tool of equilibrating pool, and the result was in concordance with the prediction by NLP transformer model. However, comparing to complex calculation, our data-driven strategy is much more convenient.

In addition, we also found an interesting phenomenon that the transformer-transfer learning model can also revise some mechanism inference work in the reported literature. For instance, Li *et al*. reported an interesting radical cascade cyclization of 1,6-enynes with aryl sulfonyl chlorides by using visible-light photoredox catalysis (Fig 10 a).[58] A possible mechanism they proposed was shown in Fig 10b. They described that an aryl radical is firstly formed by a single-electron transfer from the excited state $[Ru(bpy)_3]^{2+}$ to an aryl sulfonyl chloride, and subsequent addition of the aryl radical to the triple bond result in radical intermediate Ⅰ. Then intermediate Ⅰ underwent the cyclization reaction with the alkene to yield intermediate Ⅱ. After intramolecular cyclization process of intermediate Ⅱ, the cyclic radical was oxidized to the corresponding cyclic cation and subsequently transformed into the target product after deprotonation, accompanied with regeneration of the active $[Ru(bpy)_3]^{2+}$ species. However, our transformer-transfer learning model output another key intermediate Ⅰ & Ⅱ (Fig 10c), implying that the reaction may proceed via another pathway, which is quite different from the mechanism proposed above. With the excited $[Ru(bpy)_3]^{2+}$, sulfonyl radical is formed by the selective cleavage of the S–Cl bond.[59-60] Subsequently, sulfonyl radical adds to the terminal alkene to generate the tertiary alkyl radical. 5-exo-cyclization leads to a vinyl radical that further undergoes a 1,5-aryl migration and subsequent releasement of SO_2 to give the primary alkyl radical intermediate. The following process is as the same as Li's inference. And subsequent researches also indicated that free aryl radical is not generated in the previously reported visible light-induced cyclization of 1,6-enynes with aryl sulfony.l[61-63] Furthermore, in view of the difference in reactivity between alkyne and alkene, alkyne π-bond is stronger and less reactive than the π-bond of alkene.[50-51] Thus, the addition of sulfonyl radical into the terminal activated alkene to provide the tertiary alkyl radical is more reasonable. In this controversial mechanism inference, the mechanism directed by our model is

more reasonable, which was confirmed by Studer's investigation.[64]

Conclusion

Overall, NLP transformer model was applied to solve the task of speculating reaction mechanism, where similar cross-disciplinary studies is rarely reported. A data set of radical cascade cyclization reaction was established manually for training the model. Transfer learning was used to surmount the limitation of small data set, and the observed improvement in performance demonstrated the power of the integrated transfer learning and transformer model. The analysis of attention weights indicated that the transformer-transfer learning model exhibited a similar "thinking mode" to chemists'. Applying NLP model to the prediction of radical cascade cyclization reaction intermediates, which is convenient and effective, expands the concept of end-to-end prediction. With the accumulation of data and the continuous upgrading of algorithms, we hope that this work will empower the broader chemical community to engage with this burgeoning field and foster the growing movement of AI accelerated chemistry.

Acknowledgements

This project was supported by National Natural Science Foundation of China (No.81903438). We are grateful for the Zhejiang Provincial Key R&D Project (No. 2020C03006 & 2019-ZJ-JS-03).

Keywords: Natural Language Processing • Intermediate • Mechanism Inference • Radical Cascade Cyclization • Artificial Intelligence

[1] C. Vallance, *Nature* **2017**, *546*, 608–609.
[2] T. Deb, J. Tu, R. M. Franzini, *Chem. Rev.* **2021**, *121*, 6850–6914.
[3] M. Reiher, N. Wiebe, K. M. Svore, D. Wecker, M. Troyer, *PNAS* **2017**, 114, 7555–7560.
[4] V. Gold, *Nature* **1977**, *267*, 471–472.
[5] J. H. Chen, P. Baldi, *J. Chem. Inf. Model.* **2009**, *49*, 2034–2043.
[6] K. Jorner, T. Brinck, P. Norrby, D. Buttar, *Chem. Sci.* **2021**, *12*, 1163–1175.
[7] J. A. Keith, V. Galindo, B. Cheng, S. Chmiela, M. Gastegger, K. R. Muller, A. Tkatchenko, *Chem. Rev.* **2021**,*121*, 9816–9872.
[8] J. N. Wei, D. Duvenaud, A. Aspuru-Guzik, *ACS Cent. Sci.* **2016**, *2*, 725–732.
[9] X. Yang, Y. Wang, R. Byrne, G. Schneider, S. Yang, *Chem. Rev.* **2019**, *119*, 10520–10594.
[10] C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, K. F. Jensen, *ACS Cent. Sci.* **2017**, *3*, 434–443.
[11] A. C. Vaucher, P. Schwaller, J. Geluykens, V. H. Nair, A. Iuliano, T. Laino, *Nat. Commun.* **2021**, *12*, 2573.
[12] C. W. Coley, W. H. Green, K. F. Jensen, *Acc Chem. Res.* **2018**, *51*, 1281–1289.
[13] C. W. Coley, D. A. Thomas, J. A. M. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, R. W. Hicklin, P. P. Plehiers, J. Byington, J. S. Piotti, W. H. Green, A. J. Hart, T. F. Jamison, K. F. Jensen, *Science* **2019**, *365*, eaax1566.
[14] A. C. Mater, M. L. Coote, *J. Chem. Inf. Model* **2019**, *59*, 2545–2559.
[15] G. B. Goh, N. O. Hodas, A. Vishnu, *J. Comput. Chem.* **2017**, *38*, 1291-1307.
[16] Z. Zhang, J. A. Schott, M. Liu, H. Chen, X. Lu, B. G. Sumpter, J. Fu, S. Dai, *Angew. Chem. Int. Ed.* **2019**, *58*, 259-263.
[17] X. Qu, Y. Huang, H. Lu, T. Qiu, D. Guo, T. Agback, V. Orekhov, Z. Chen, *Angew. Chem. Int. Ed.* **2020**, *59*, 10297-10300.
[18] H. Li, *Natl. Sci. Rev.* **2018**, *5*, 22–24.
[19] J. Kim. J. Nam, *arXiv: 1612.09529v1* **2016**.
[20] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, A. A. Lee, *ACS Cent. Sci.* **2019**, *5*, 1572–1583.
[21] S. Steiner, J. Wolf, S. Glatzel, A. Andreou, J. M. Granda, G. Keenan, T. Hinkley, G. Aragon-Camarasa, P. J. Kitson, D. Angelone, L. Cronin, *Science* **2019**, *363*, eaav2211.
[22] E. M. Gale, D. J. Durand, *Nat. Chem.* **2020**, *12*, 509–510.
[23] B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. Luu Nguyen, S. Ho, J. Sloane, P. Wender, V. Pande, *ACS Cent. Sci.* **2017**, *3*, 1103–1113.
[24] H. Duan, L. Wang, C. Zhang, L. Guo, J. Li, *RSC Advances* **2020**, *10*, 1371–1378.
[25] K. Lin, Y. Xu, J. Pei, L. Lai, *Chem. Sci.* **2020**, *11*, 3355–3364.
[26] P. Schwaller, R. Petraglia, V. Zullo, V. H. Nair, R. A. Haeuselmann, R. Pisoni, C. Bekas, A. Iuliano, T. Laino, *Chem. Sci.* **2020**, *11*, 3316–3325.
[27] I. V. Tetko, P. Karpov, R. Van Deursen, G. Godin, *Nat. Commun.* **2020**, *11*, 5575.
[28] J. Xuan, A. Studer, *Chem. Soc. Rev.* **2017**, *46*, 4329–4346.
[29] J. Liao, X. Yang, L. Ouyang, Y. Lai, J. Huang, R. Luo, *Org. Chem. Front.* **2021**, *8*, 1345–1363.
[30] X. Y. Liu, Y. Qin, *Acc. Chem. Res.* **2019**, *52*, 1877–1891.
[31] J. K. Qiu, B. Jiang, Y. L. Zhu, W. J. Hao, D. C. Wang, J. Sun, P. Wei, S. J. Tu, G. Li, *J. Am. Chem. Soc.* **2015**, *137*, 8928–8931.
[32] Y. Li, B. Liu, R.-J. Song, Q.-A. Wang, J.-H. Li, *Adv. Synth. Catal.* **2016**, *358*, 1219–1228.
[33] S. B. Sheng-ze Zhou, and John A. Murphy, *Org. Lett.* **2002**, *4*, 443–445.
[34] https://pubchem.ncbi.nlm.nih.gov
[35] https://www.ebi.ac.uk/chembl/
[36] https://www.reaxys.com/
[37] https://scifinder.cas.org
[38] H. Ozturk, A. Ozgur, P. Schwaller, T. Laino, E. Ozkirimli, *Drug Discov Today* **2020**, *25*, 689–705.
[39] Y. Zhang, L. Wang, X. Wang, C. Zhang, J. Ge, J. Tang, A. Su, H. Duan, *Org. Chem. Front.* **2021**, *8*, 1415–1423.
[40] R. Bai, C. Zhang, L. Wang, C. Yao, J. Ge, H. Duan, *Molecules* **2020**, *25*, 2357–2371.
[41] L. Wang, C. Zhang, R. Bai, J. Li, H. Duan, *Chem. Commun.* **2020**, *56*, 9368–9371.
[42] Y. Wu, C. Zhang, L. Wang, H. Duan, *Chem. Commun.* **2021**, *57*, 4114–4117.
[43] D. M. Lowe, Chemical reactions from US patents (1976-Sep2016); https://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873, **2017**.
[44] G. Pesciullesi, P. Schwaller, T. Laino, J. L. Reymond, *Nat. Commun.* **2020**, *11*, 4874.
[45] D. Kreutter, P. Schwaller, J. L. Reymond, *Chem. Sci.* **2021**, *12*, 8648–8659.
[46] Z. Ding, M. Shao, Y. Fu, *IEEE Trans Neural Netw Learn Syst* **2018**, *29*, 310–323.
[47] J. Wang, D. Agarwal, M. Huang, G. Hu, Z. Zhou, C. Ye, N. R. Zhang, *Nat. Methods* **2019**, *16*, 875–878.
[48] T. Foldes, A. Madarasz, A. Revesz, Z. Dobi, S. Varga, A. Hamza, P. R. Nagy, P. M. Pihko, I. Papai, *J. Am. Chem. Soc.* **2017**, *139*, 17052–17063.
[49] R. Fu, W. Hao, Y. Wu, N. Wang, S. Tu, G. Li, B. Jiang, *Org. Chem. Front.* **2016**, *3*, 1452–1456.
[50] R. K. Mohamed, S. Mondal, B. Gold, C. J. Evoniuk, T. Banerjee, K. Hanson, I. V. Alabugin, *J. Am. Chem. Soc.* **2015**, *137*, 6335–6349.
[51] I. V. Alabugin, E. Gonzalez-Rodriguez, *Acc. Chem. Res.* **2018**, *51*, 1206–1219.
[52] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G. Z. Yang, *Sci. Robot* **2019**, *4*, eaay7120.
[53] J. Stoyanovich, J. J. Van Bavel, T. V. West, *Nat. Mach. Intell.* **2020**, *2*, 197–199.
[54] L. Kohoutova, J. Heo, S. Cha, S. Lee, T. Moon, T. D. Wager, C. W. Woo, *Nat. Protoc.* **2020**, *15*, 1399–1435.
[55] D. P. Kovacs, W. McCorkindale, A. A. Lee, *Nat. Commun.* **2021**, *12*, 1695.
[56] P. Schwaller, T. Gaudin, D. Lanyi, C. Bekas, T. Laino, *Chem. Sci.* **2018**,

9, 6091–6098.
[57] R. K. M. Sayantan Mondal, M. Manoharan, H. Phan, and I. V. Alabugin, *Org. Lett.* **2013**, *15*, 5650–5653.
[58] G. B. Deng, Z. Q. Wang, J. D. Xia, P. C. Qian, R. J. Song, M. Hu, L. B. Gong, J. H. Li, *Angew. Chem. Int. Ed.* **2013**, *52*, 1535–1538.
[59] C. J. Wallentin, J. D. Nguyen, P. Finkbeiner, C. R. Stephenson, *J. Am. Chem. Soc.* **2012**, *134*, 8875–8884.
[60] H. Jiang, X. Chen, Y. Zhang, S. Yu, *Adv. Synth. Catal.* **2013**, *355*, 809–813.
[61] X. Meng, Q. Kang, J. Zhang, Q. Li, W. Wei, W. He, *Green Chem.* **2020**, *22*, 1388–1392.
[62] T. H. Zhu, X. C. Zhang, X. L. Cui, Z. Y. Zhang, H. Jiang, S. S. Sun, L. L. Zhao, K. Zhao, T. P. Loh, *Adv. Synth. Catal.* **2019**, *361*, 3593–3598.
[63] L. Chen, M. Zhou, L. Shen, X. He, X. Li, X. Zhang, Z. Lian, *Org. Lett.* **2021**, *23*, 4991–4996.
[64] J. Xuan, D. Gonzalez-Abradelo, C. A. Strassert, C.-G. Daniliuc, A. Studer, *Eur. J. Org. Chem.* **2016**, *29*, 4961–4964.