Overview of Synthesis Planning and Requirements for Machine Learning Approaches

Ernesto Gomez

MCSC UROP

Introduction

In chemistry, synthesis planning is the process of generating a plan or sequence of chemical reactions to produce a specific target molecule. The goal of synthesis planning is to find the most efficient and cost-effective route to the target molecule, taking into account factors such as reactivity, availability of materials, and potential side reactions. Pioneered by E. J. Corey in the 1960s, retrosynthetic analysis through a disconnection approach introduced the development of reusable protocols that could be applied when designing a synthetic plan.[1] Corey offered the first concrete algorithm for producing a logical synthesis of a target molecule. The technique involves reducing the target molecule into a sequence of progressively simpler structures along a pathway which ultimately leads to the identification of simple or commercially available starting materials (example shown in Figure 1). When done manually, retrosynthetic analysis produces routes that depend on the knowledge of the chemist, their inherent biases toward specific reactions, and the equipment available to them. It is common to find "conversations" in the literature where one chemist developed a synthetic route to a target molecule and publishes the route in an article. Afterward, another chemist chooses to replace one or more steps of the previous author with their preferences and reports the benefits/drawbacks.[2,3] Such "conversations" are excellent examples of ongoing peer review in science to evolve human knowledge. However, the process can take many years for a single iteration, making it slow and very expensive.

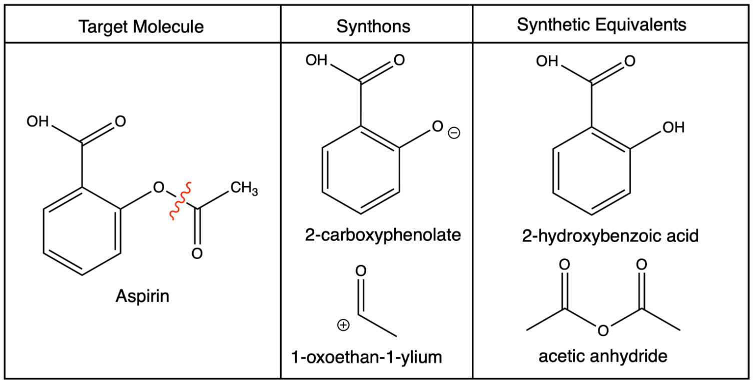


Figure 1: An example of retrosynthesis in which Aspirin is the target molecule and precursors, or synthetic equivalents to the synthons, are found to form Aspirin.

Since the 1960s, various computer-aided synthesis planning (CASP) applications have been developed to mimic chemists' thinking and help organic chemists in their work, based on the principles established by Corey.[4–12] Ideally, these CASP tools would be given a target molecule input, and work backward to generate a list of reactions that each connect that target molecule to commercially available starting materials through a series of chemically attainable reaction steps. CASP tools generally consist of five major components: A template library containing rules by which disconnections are proposed; A recursive template application engine that generates candidate reactants for target product molecules; a database containing commercially available compounds; a strategy to guide the retrosynthetic search toward chemicals in that database; A method for single-step or pathway-level scoring, creating preference for fewer steps.[13]

Given the size and complexity of chemical space, the evolution of CASP applications has come to include machine learning techniques.[13–16] Such techniques can approximate complex functions where the exact relationship between input and output is not easily expressed. Increased performance in machine learning techniques can be generally attributed to a combination of improvements in hardware capability and data availability. As it stands, chemical databases and automated management of synthetic knowledge such as Reaxys[17], SciFinder[18], ChemSpider[19], and SPRESI[20] are invaluable tools to organic chemists, especially for those searching for literature sources or examples of analogous reactions. Searches are mostly manual, step-by-step procedures with only rudimentary capabilities to evaluate reaction sequences, which underuses the growing power of the modern computer.[11]

Deep learning, a type of machine learning, has proved itself to be a promising direction for retrosynthesis planning. In 2017, Seglers et al.[21] combined the strengths of neural networks and symbolic reasoning to enable this novel machine learning approach for retrosynthesis and reaction prediction. Deep learning improved the accuracy of predictions and the interpretability of the models, making it a significant advance in the field. The success of deep learning can be attributed to its nature of having: automated feature extraction, which replaces the need for manual feature engineering; the capacity to handle large complex data; and generalizing learned patterns to new cases.[22]

Data Sourcing

The construction of deep learning algorithms, similar to traditional statistical methods, requires sufficient data to be supplied for model training. Sources for chemical reaction data across institutions and organizations differ in scope, format applicability, authorization quality, and other aspects.[23] Frequently used datasets include Reaxys[24], SciFinder[25,26], the United States Patent and Trademark Office (USPTO), and Pistachio.

Reaxys, operated by Elsevier, offers users an interface and database to retrieve relevant chemical literature, patent information, valid compound properties, and experimental procedures. Reagent information, reaction conditions, and yields are provided as non-standardized text as they are collected from publications. For this reason, reaction data from Reaxys requires additional data processing to standardize them and make them useful for any machine learning purposes.

Similar to Reaxys, SciFinder, produced by Chemical Abstracts Service (CAS), is a much more comprehensive database containing chemical literature including patents, journals, conference papers, and web resources.[27]

Given the cost of Reaxys and SciFinder, researchers may look to the open-source United States Patent and Trademark Office (USPTO). Due to the size of the USPTO database, extracted portions of the USPTO are used to train and validate as separate databases. The USPTO dataset became USPTO-50k and USPTO-full as subsets of preprocessed chemical reactions.[28] The popular USPTO-50k contains 50,000 randomly selected reactions that were classified into 10 reaction classes.

Alternatively, researchers look to Pistachio, supported by NextMove Software.[29] Pistachio experiences continuous updates to its data mining pipeline and adds newly patented reactions to the dataset. As a result, Pistachio is a superset of the USPTO in addition to the European Patent Office (EPO), and the World Intellectual Property Organization (WIPO).

Data Representations

Chemical structures and chemical structure transformations (reactions) are at the core of cheminformatics. The efficiency of cheminformatics applications, such as deep-learning-based retrosynthesis, is tightly coupled with the adequate representation of the structure and reaction of molecules. [30] As a result, the quality of the machine-readable data representations of the chemical reaction will directly affect the subsequent application.

One approach to modeling chemical reactions is using a graphical representation. Condensed graphs of reaction (CGR) are a popular cheminformatics reaction representation. The CGR is a superposition of the reactant and product graphs of a chemical reaction and thus an ideal input for graph-based machine learning approaches.[31] A CGR (figure 2) contains all atoms involved in a reaction as vertices (or nodes) connected by all

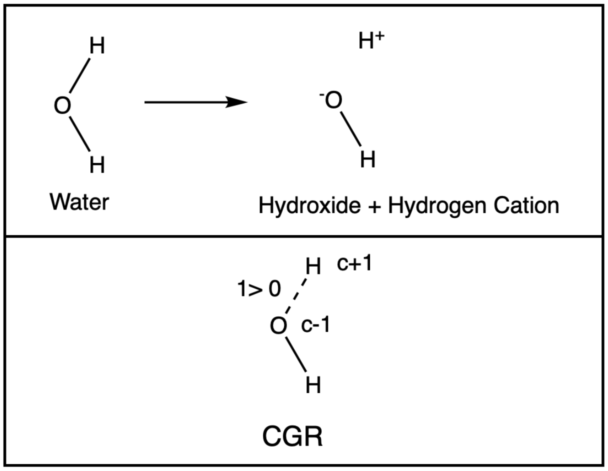


Figure 2: The reaction schema showing dissociation of water (top) and equivalent condensed graph of reaction (bottom).

bonds (formed, created, and broken) as edges (or links). This graphical representation indicates changes in the properties of atoms and bonds. This approach does require atom-to-atom mapping, such that the behavior of each atom and bond in the reaction is accounted for properly.[32] CGRTools is a tool for processing reactions based on the CGR approach. For more details on how CGRTools operates, the reader is directed to further reading.[33]

In addition to graph representations, researchers have developed various machine-readable string notations for representing chemical reactions. Reaction SMILES (Simplified Molecular Input Line Entry System) is a linear notation for describing chemical reactions similar to SMILES. SMILES uses a string to describe molecular structures, while reaction SMILES (figure 3) represent each reactant and product with a SMILES string. The reaction itself is represented by the symbol ">", following the form, [reactants]>[agents]>[products].[34] While SMILES is a powerful and flexible approach to describing reactions, different researchers studying the same reaction may generate different descriptions of one reaction. As a result, the RInChI (International Chemical Identifier for Reactions) project, an extension of InChI (International Chemical Identifier), set out to create unambiguous descriptions for reactions.[35] RInChI grammar, however, is relatively more complicated than that of Reaction SMILES.[36]

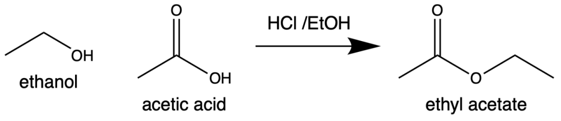
Reaction Scheme	 OH ethanol O OH acetic acid HCl /EtOH O O ethyl acetate
Reaction SMILES	CC(=O)O.OCC>[H+].[Cl-].OCC>CC(=O)OCC

Figure 3: Example of reaction SMILES and equivalent reaction Scheme

Atom Mapping

Inconsistent data quality remains an issue to be addressed in the data preparation stage. Atom-to-atom mapping (AAM) often ranges from poor to nonexistent.[37] AAM refers to the application of a procedure that establishes a correspondence between the atoms of reactants and products. AAM assists in the identification of a reaction center which aids in the preparation of reaction templates, which are to be discussed.[38] Accurate atom-atom mapping can facilitate downstream tasks such as calculating the number of conserved carbon atoms in a reaction to determine the metabolic efficiency or tracking atoms to understand and demonstrate the reaction mechanism.[39] Attempts to develop accurate atom-to-atom mapped reactions reflect a larger issue of unknown reaction mechanisms, which often require a deeper understanding of chemical reactions than available in current scientific literature. Methods of AAM include structural and optimization-based approaches.

Structure-based approaches determine common chemical structures to match the atoms of the reactant with that of the product. Tools such as AutoMapper[40] and ICMAP[41] pioneer this method. AutoMapper accepts many chemical formats such as InChI and SMILES but cannot identify the reaction center. Edit with ZoteroICMAP, however, can mark the reaction center. Structure-based approaches, however, function on the condition that the reactants and products maintain similar enough substructure to be compared and mapped.

Optimization-based approaches use Mixed Integer Linear Optimization (MILP) to identify atom mappings.[40] Determination of REAction Mechanisms (DREAM)[39] and Minimum Weighted Edit-Distance (MWED)[38] are tools that use such an approach. Optimization-based approaches aim to minimize the number of bonds broken, bonds formed, and bond order changes, between reactants and products. Like DREAM, MWED assigns weights to bonds of the molecules in the reaction and a specific cost when a bond is modified.[42]

Machine Learning Models

Once reaction data has been formatted and if necessary, mapped, it can be fed into a machine-learning model to be trained. Methods for constructing machine learning models of retrosynthesis planning can be grouped into template-based and template-free. Template-free methods include those that are sequence-based and those that are graph-based.

The template-based approach compares the target molecule with a large set of templates to determine potential precursors. The template is a set of reaction rules that consist of a set of minimal transformations to characterize a chemical reaction.[43] These templates are extracted from atom-mapped reaction examples.[13] The purpose of using templates is to find a connection from this target molecule to potential precursors. A problem arises, however, when it comes to selecting the appropriate reaction template. Accounting for functional groups is another problem that is addressed appropriately by Segler et al.[44] such that molecular context is accounted for in their modeling. While template-based methods are highly interpretable there are two common problems: poor generalization as the search of precursors is limited to extracted templates; and poor scalability as the number of candidate precursors increases[13].

The template-free approach finds unclear relationships about reaction mechanisms in data. Sequence-based methods, a template-free approach, use neural sequence-to-sequence models that learn from patent data to perform retrosynthetic reaction prediction. The model is trained end-to-end eliminating the need for reaction rules and atom mapping. Graph-based methods, another template-free approach, use graph neural networks (GNN) that can be taught through condensed graphs of reactions (CGR)[31]. The goal of GNNs is to learn the representations of each atom by aggregating representations of its neighbors through messages passing across the molecular graph. The learned representations can be used to predict the reaction properties of the molecule.[31,33,45,45]

Conclusion and Outlook

Deep learning-based methods for retrosynthesis planning are being applied in organic synthesis and drug discovery, with the potential to advance personalized medicine. Throughout the analysis of contemporary literature, there is an apparent lack of data related to conditions fed into deep learning models such as reagents, catalysis, solvents, byproducts, and temperature. A potential issue with the inclusion of this data, and upscaling of current methods, include computational difficulties. To overcome these computational challenges, there must either be an increase in hardware capabilities or improvements in current methods such as improvement on focused pathway predictors such as those that use the Monte Carlo Tree Search to quickly sample precursors and resultingly relieve computational demands.[21] Additionally, an improvement of current data sources would greatly streamline the entire process. While the USPTO is a very popular dataset for model training, it would be appropriate to construct a standardized dataset that would provide deep learning algorithms with sufficient training data in a readily available manner. Ultimately, a more perfect retrosynthesis algorithm would require continued collaboration between chemists and computer scientists to be developed. Especially when working in deep learning architectures in such a chemically rich context.

(1) Corey, E. J. General Methods for the Construction of Complex Molecules. **1967**, *14* (1), 19–38. https://doi.org/10.1351/pac196714010019.

(2) Lavallo, V.; Canac, Y.; Präsang, C.; Donnadieu, B.; Bertrand, G. Stable Cyclic (Alkyl)(Amino)Carbenes as Rigid or Flexible, Bulky, Electron-Rich Ligands for Transition-Metal Catalysts: A Quaternary Carbon Atom Makes the Difference. *Angew. Chem. Int. Ed.* **2005**, *44* (35), 5705–5709. https://doi.org/10.1002/anie.200501841.

(3) Jazzar, R.; Dewhurst, R. D.; Bourg, J.-B.; Donnadieu, B.; Canac, Y.; Bertrand, G. Intramolecular "Hydroiminiumation" of Alkenes: Application to the Synthesis of Conjugate Acids of Cyclic Alkyl Amino Carbenes (CAACs). *Angew. Chem. Int. Ed.* **2007**, *46* (16), 2899–2902. https://doi.org/10.1002/anie.200605083.

(4) Corey, E. J.; Wipke, W. T. Computer-Assisted Design of Complex Organic Syntheses. *Science* **1969**, *166* (3902), 178–192.

(5) Corey, E. J.; Wipke, W. T.; Cramer, R. D.; Howe, W. J. Computer-Assisted Synthetic Analysis. Facile Man-Machine Communication of Chemical Structure by Interactive Computer Graphics. *J. Am. Chem. Soc.* **1972**, *94* (2), 421–430. https://doi.org/10.1021/ja00757a020.

(6) Wipke, W. T.; Ouchi, G. I.; Krishnan, S. Simulation and Evaluation of Chemical Synthesis—SECS: An Application of Artificial Intelligence Techniques. *Artif. Intell.* **1978**, *11* (1), 173–193. https://doi.org/10.1016/0004-3702(78)90016-4.

(7) Hendrickson, J. B.; Toczko, A. G. SYNGEN Program for Synthesis Design: Basic Computing Techniques. *J. Chem. Inf. Model.* **1989**, *29* (3), 137–145. https://doi.org/10.1021/ci00063a001.

(8) Ihlenfeldt, W.-D.; Gasteiger, J. Computer-Assisted Planning of Organic Syntheses: The Second Generation of Programs. *Angew. Chem. Int. Ed. Engl.* **1996**, *34* (23–24), 2613–2633. https://doi.org/10.1002/anie.199526131.

(9) Satoh, K.; Funatsu, K. A Novel Approach to Retrosynthetic Analysis Using Knowledge Bases Derived from Reaction Databases. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (2), 316–325. https://doi.org/10.1021/ci980147y.

(10) Law, J.; Zsoldos, Z.; Simon, A.; Reid, D.; Liu, Y.; Khew, S. Y.; Johnson, A. P.; Major, S.; Wade, R. A.; Ando, H. Y. Route Designer: A Retrosynthetic Analysis Tool Utilizing Automated Retrosynthetic Rule Generation. *J. Chem. Inf. Model.* **2009**, *49* (3), 593–602. https://doi.org/10.1021/ci800228y.

(11) *Computer‐Assisted Synthetic Planning: The End of the Beginning - Szymkuć - 2016 - Angewandte Chemie International Edition - Wiley Online Library*. https://onlinelibrary.wiley.com/doi/full/10.1002/anie.201506101 (accessed 2023-01-26).

(12) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555* (7698), 604–610. https://doi.org/10.1038/nature25978.

(13) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51* (5), 1281–1289. https://doi.org/10.1021/acs.accounts.8b00087.

(14) Aldeghi, M.; Coley, C. W. A Focus on Simulation and Machine Learning as Complementary Tools for Chemical Space Navigation. *Chem. Sci.* **2022**, *13* (28), 8221–8223. https://doi.org/10.1039/D2SC90130G.

(15) Wang, X.; Qian, Y.; Gao, H.; Coley, C. W.; Mo, Y.; Barzilay, R.; Jensen, K. F. Towards Efficient Discovery of Green Synthetic Pathways with Monte Carlo Tree Search and Reinforcement Learning. *Chem. Sci.* **2020**, *11* (40), 10959–10972. https://doi.org/10.1039/D0SC04184J.

(16) Soleimany, A. P.; Amini, A.; Goldman, S.; Rus, D.; Bhatia, S. N.; Coley, C. W. Evidential Deep Learning for Guided Molecular Property Prediction and Discovery. *ACS Cent. Sci.* **2021**, *7* (8), 1356–1367. https://doi.org/10.1021/acscentsci.1c00546.

(17) *Reaxys - An expert-curated chemistry database*. https://www.elsevier.com/solutions/reaxys (accessed 2023-01-26).

(18) *CAS Analytical Methods*. CAS. https://www.cas.org/solutions/cas-scifinder-discovery-platform/cas-analytical-methods (accessed 2023-01-26).

(19) *ChemSpider | Search and share chemistry*. http://www.chemspider.com/ (accessed 2023-01-26).

(20) *SPRESIweb - chemical structure and reaction database by InfoChem*. https://www.spresi.com/ (accessed 2023-01-26).

(21) Segler, M.; Preuß, M.; Waller, M. P. Towards “AlphaChem”: Chemical Synthesis Planning with Tree Search and Deep Neural Network Policies. arXiv January 31, 2017. http://arxiv.org/abs/1702.00020 (accessed 2023-01-29).

(22) Schulz, H.; Behnke, S. Deep Learning. *KI - Künstl. Intell.* **2012**, *26* (4), 357.

(23) Dong, J.; Zhao, M.; Liu, Y.; Su, Y.; Zeng, X. Deep Learning in Retrosynthesis Planning: Datasets, Models and Tools. *Brief. Bioinform.* **2022**, *23* (1), bbab391. https://doi.org/10.1093/bib/bbab391.

(24) Goodman, J. Computer Software Review: Reaxys. *J. Chem. Inf. Model.* **2009**, *49* (12), 2897–2898. https://doi.org/10.1021/ci900437n.

(25) *Information Retrieval: SciFinder, 2nd Edition | Wiley*. Wiley.com. https://www.wiley.com/en-us/Information+Retrieval%3A+SciFinder%2C+2nd+Edition-p-9780470749425 (accessed 2023-01-30).

(26) Meloche, K. J.; Mears, J.; Schenck, R. J. Intriguing Records in CAS Databases. In *A Festival of Chemistry Entertainments*; ACS Symposium Series; American Chemical Society, 2013; Vol. 1153, pp 21–40. https://doi.org/10.1021/bk-2013-1153.ch002.

(27) Menon, A.; Krdzavac, N. B.; Kraft, M. From Database to Knowledge Graph — Using Data in Chemistry. *Curr. Opin. Chem. Eng.* **2019**, *26*, 33–37. https://doi.org/10.1016/j.coche.2019.08.004.

(28) Schneider, N.; Stiefl, N.; Landrum, G. A. What’s What: The (Nearly) Definitive Guide to Reaction Role Assignment. *J. Chem. Inf. Model.* **2016**, *56* (12), 2336–2346. https://doi.org/10.1021/acs.jcim.6b00564.

(29) *NextMove Software | Pistachio*. https://www.nextmovesoftware.com/pistachio.html (accessed 2023-02-08).

(30) Kochev, N.; Avramova, S.; Jeliazkova, N. Ambit-SMIRKS: A Software Module for Reaction Representation, Reaction Search and Structure Transformation. *J. Cheminformatics* **2018**, *10* (1), 42. https://doi.org/10.1186/s13321-018-0295-6.

(31) Heid, E.; Green, W. H. Machine Learning of Reaction Properties via Learned Representations of the Condensed Graph of Reaction. *J. Chem. Inf. Model.* **2022**, *62* (9), 2101–2110. https://doi.org/10.1021/acs.jcim.1c00975.

(32) Polishchuk, P.; Madzhidov, T.; Gimadiev, T.; Bodrov, A.; Nugmanov, R.; Varnek, A. Structure-Reactivity Modeling Using Mixture-Based Representation of Chemical Reactions. *J. Comput. Aided Mol. Des.* **2017**, *31* (9), 829–839. https://doi.org/10.1007/s10822-017-0044-3.

(33) 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. https://doi.org/10.1021/acs.jcim.9b00102.

(34) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36. https://doi.org/10.1021/ci00057a005.

(35) Grethe, G.; Goodman, J. M.; Allen, C. H. International Chemical Identifier for Reactions (RInChI). *J. Cheminformatics* **2013**, *5* (1), 45. https://doi.org/10.1186/1758-2946-5-45.

(36) David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. Molecular Representations in AI-Driven Drug Discovery: A Review and Practical Guide. *J. Cheminformatics* **2020**, *12* (1), 56. https://doi.org/10.1186/s13321-020-00460-5.

(37) Fooshee, D.; Andronico, A.; Baldi, P. ReactionMap: An Efficient Atom-Mapping Algorithm for Chemical Reactions. *J. Chem. Inf. Model.* **2013**, *53* (11), 2812–2819. https://doi.org/10.1021/ci400326p.

(38) Lin, A.; Dyubankova, N.; Madzhidov, T. I.; Nugmanov, R. I.; Verhoeven, J.; Gimadiev, T. R.; Afonina, V. A.; Ibragimova, Z.; Rakhimbekova, A.; Sidorov, P.; Gedich, A.; Suleymanov, R.; Mukhametgaleev, R.; Wegner, J.; Ceulemans, H.; Varnek, A. Atom-to-Atom Mapping: A Benchmarking Study of Popular Mapping Algorithms and Consensus Strategies. *Mol. Inform.* **2022**, *41* (4), 2100138. https://doi.org/10.1002/minf.202100138.

(39) Latendresse, M.; Malerich, J. P.; Travers, M.; Karp, P. D. Accurate Atom-Mapping Computation for Biochemical Reactions. *J. Chem. Inf. Model.* **2012**, *52* (11), 2970–2982. https://doi.org/10.1021/ci3002217.

(40) Jaworski, W.; Szymkuc, S.; Mikulak-Klucznik, B.; Piecuch, K.; Klucznik, T.; Kazmierowski, M.; Rydzewski, J.; Gambin, A.; Grzybowski, B. A. Automatic mapping of atoms across both simple and complex chemical reactions. **2019**.

(41) Kraut, H.; Eiblmaier, J.; Grethe, G.; Löw, P.; Matuszczyk, H.; Saller, H. Algorithm for Reaction Classification. *J. Chem. Inf. Model.* **2013**, *53* (11), 2884–2895. https://doi.org/10.1021/ci400442f.

(42) Preciat Gonzalez, G. A.; El Assal, L. R. P.; Noronha, A.; Thiele, I.; Haraldsdóttir, H. S.; Fleming, R. M. T. Comparative Evaluation of Atom Mapping Algorithms for Balanced Metabolic Reactions: Application to Recon 3D. *J. Cheminformatics* **2017**, *9* (1), 39. https://doi.org/10.1186/s13321-017-0223-1.

(43) Baylon, J. L.; Cilfone, N. A.; Gulcher, J. R.; Chittenden, T. W. Enhancing Retrosynthetic Reaction Prediction with Deep Learning Using Multiscale Reaction Classification. *J. Chem. Inf. Model.* **2019**, *59* (2), 673–688. https://doi.org/10.1021/acs.jcim.8b00801.

(44) Segler, M. H. S.; Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. – Eur. J.* **2017**, *23* (25), 5966–5971. https://doi.org/10.1002/chem.201605499.

(45) Dai, H.; Li, C.; Coley, C.; Dai, B.; Song, L. Retrosynthesis Prediction with Conditional Graph Logic Network. In *Advances in Neural Information Processing Systems*; Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F. d’, Fox, E., Garnett, R., Eds.; Curran Associates, Inc., 2019; Vol. 32.