Machine learning in computer-aided synthesis planning

Ximing Li
Department of Chemistry, Princeton University
ximingl@princeton.edu

Abstract

Although the capability to challenge humans in various strategic games has been widely demonstrated, the use of machine learning in the automated planning of organic syntheses remains unprecedented. As the huge impact that such a tool, if feasible, could have on the synthetic community, an experiment is shown in this project proposal where a machine learning algorithm could design syntheses leading to unprecedented synthetic target by introducing reasonable retrosynthesis. An example of this idea is mainly shown by discussing a single stereotyped chemistry due to the complexity and immenseness of the whole chemistry reaction library.

1 Introduction

An artificial synthetic organic design algorithm, which will automatically construct and manipulate robust and viable synthetic routes based on current human knowledge of organic chemistry reactivity, is a notably challenging machine learning problem without direct amenability to existing approaches.

Expectably, equipping computers with strong, robust learning algorithms, symbolic representations, and the access to the rich history of human-developed reactions, will probably also open the door for machines to becoming masters in this discipline, which has happened recently in many other areas such as man-machine champion games (chess, Go, and some video games).

In this work, the idea of computer-aided synthesis planning will be developed towards the goal of machines auto-learning from database and judging the most optimal methodology sequence for a given organic synthesis target. Ideally, the algorithm which this machine will be equipped with should be able to directly learn from the readily available online reaction database (SciFinder, Reaxys, Patent Office Database) which filled with reaction examples (rather than from expert-biased so-called reaction rules), read the target molecule in a symbolic representative and chemically meaningful way, and generate a library of ranked possible reaction sequences which will fulfill the given task.

2 Part I. Defining Reactions

In a typical reaction database, a chemical structure could be presented as an alphanumeric string ("SMILES" structure) for the convenience of searching the reactions of it as a substructure. For the purpose of feature selection, I would like to setup stereotyped reactions (total number might 1,000 as an estimated number for all known organic chemistry reaction to human being) to fulfill the same purpose. A given reaction could be classified as one of these stereotyped reactions, and then featured by the answer to a serious of questions from a decision tree, which are ranked in terms of importance to the of this type chemistry. For example, an ideal decision tree for one of the reactions rules (condensation of esters with aldehydes) would be like this: (Figure 1)

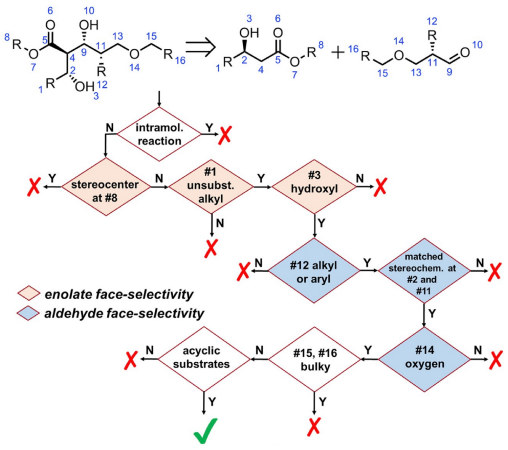

Figure 1: Defining Reactions

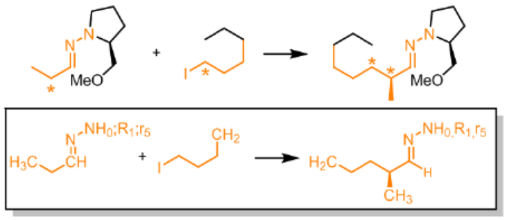

Figure 2: Mechanistic Rules(1)

The tree begins with a condition of the reaction being intermolecular, which is the key question chemists want to ask for this type of chemistry. To ensure face selectivity of the enolate, conditions for the substituents at positions #8, #1, and #3 are also considered important, thus being asked in order. Conditions at positions #12, #2, #11, #14 follow, for ensuring proper face selectivity of the aldehyde. For the last two questions, substrates should be acyclic because cyclic structures might distort the aldehyde-titanium chelate conformation or face selectivity of the ester enolate. The other requirement concerns the consonant selectivity at both substrates that ensures the desired diastereoselectivity. Once the answers to these questions are known, a chemical reaction could ideally be turned into a string of numbers by their features.

3 Part II. Mechanistic Rules

For the purpose of understanding what exactly happens in a given chemical reaction, it is sometimes necessary to extract the "reaction core", as a matter of fact typically not every element of a substrate is participating in the reaction. As the feature selection has been processed, this process could be simplified. Here demonstrated an example of a literature-reported transformation from which the "reaction core" is extracted: The core is colored in orange, covers atoms changing their local environments, and also includes flanking atoms up to three bonds away. However, even with this extended neighborhood, the transform does not capture the influence of a distant stereo-directing group, CH2OMe, which would lead to a problem later. (Figure 2)

Subsequently, application of this automatically-extracted reaction core to various synthetic targets could classify most of the reactions that we're potentially interested in. However, it could be complicated under certain circumstances. Here demonstrated a case where reaction rule extracted successfully applied to Epothilone A intermediate (top) but met with a problem with the substrate with a nitropentane side chain (middle). In the latter case, the reaction is not feasible since the pendant nitroalkyl group is incompatible with lithiated azaenolate formed from the hydrazone upon the initial treatment with LDA, thus disabled the reaction. Another case emerged where in the absence of the distant stereo-directing group, which is also not included in the reaction core, the transform

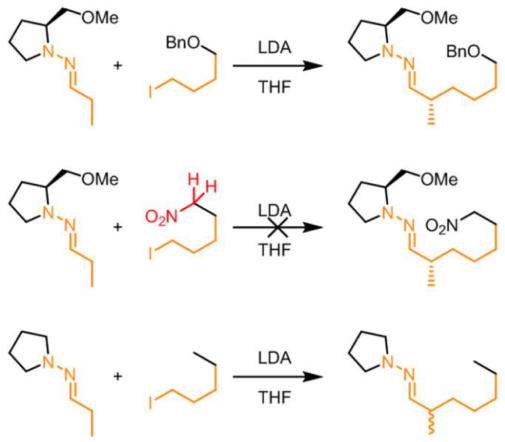

Figure 3: Mechanistic Rules(2)

may still try to predict stereoselective outcome, but as a matter of fact a racemic mixture will be obtained in experiments (bottom). For solving this problem, a great number of manually given rules (assembly over 10,000) is required to increase the accuracy of prediction.(Figure 3)

4 Part III. Chemistry beyond Records

After the basic reaction identification, classification, core extraction and manual addition of mechanistic rules, most of the reactions could be potentially predicted in a reasonably manner. However, there are still a great amount of unprecedented reactions, or reactions evolving complicated structures which contain multiple possible reaction sites. Naturally, one could argue that this selection could be operated by QM calculation, which is a well-developed mature method to calculate the most reactive position within a molecule towards multiples reactions. However, such a method may take huge amount of processing time considered the amount of candidate reaction process generated by the codes. To overcome some these limitations, an additional empirical/literature-result-based selection process that includes several known reaction rules predictions is considered. For example, in the following example of predicting the most active atom for electrophilic aromatic substitution (EAS), input molecule is divided into separate single ring system. Subsequently, depending on the ring type, the most active position in each ring is determined using Hammett-based model or more elaborate heuristics. Subsequent removal of less active rings begins with the analysis of the fused system. (Figure 4)

5 Part IV. Exploring for Complete Pathways

Having the above rules for individual synthetic moves set up, the searching and exploring of complete synthetic pathways is the critical function of this work. Trying to exhaust all the possible routes is perhaps impractical. Imagine there are an estimate order of 10 possible synthetic moves the computer needs to consider at each synthetic step, then for a common 10-step target synthesis, there are $10\hat{1}0$ possible routes leading to the desired target, way too large to explore in an exhaustive fashion. The only way to avoid this complication is to teach the machine to search the possibilities in an efficient manner, by introducing the scoring function evaluating the "synthetic value" of each reaction at each layer. Obviously, if the same synthon is reached by 1) two reactions from the target vs ten reactions, or, 2) a reaction with two precedented similar reactions vs ten precedents, or, 3) a sequence with 45% overall yield vs 15% yield, the shorter, more precedented, more efficient solution should have a better score. Ideally, the use of this function would enable us to filter most of the synthetic choices off to create convenient walk over the enormous synthetic graphs, and ultimately identifying and ranking the best-scoring solutions.

The exploration pathway is similar to the shortest path search algorithm. First a layer of "moves" is generated to be further retro-synthesized/expanded. Such an expansion layer is added to the first one,

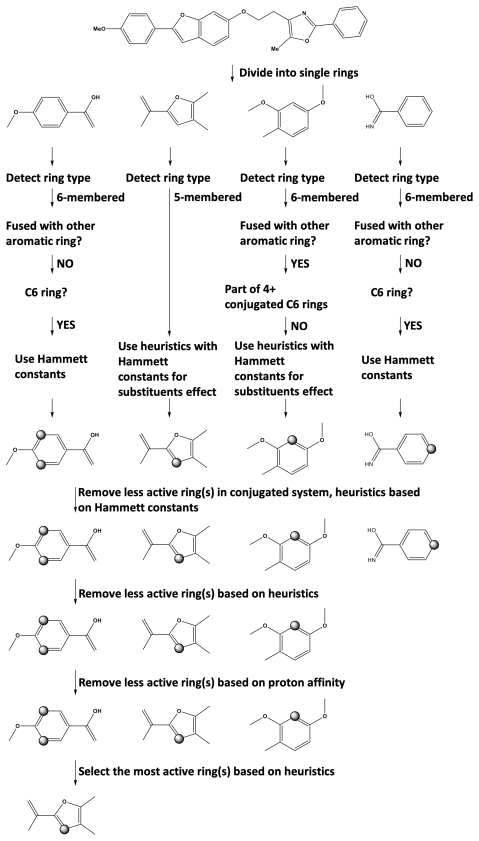

Figure 4: Chemistry beyond Records

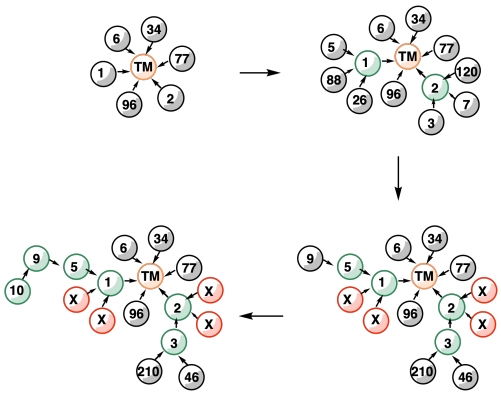

Figure 5: Exploring for Complete Pathways

and new choices are immediately made based on scoring function. In such way the most efficient pathway can be quickly located through the generation of layers of possible operations. Here is a scheme illustrating exploration of the synthesis graph: (1) Originally, the current layer explored consists of only the target and the first-generation moves surrounded. All these incoming moves are calculated by scoring function, with a given values are 1, 2, ... (2) The currently lowest-scoring move (1), (2) is further expanded with the subsequent layer calculated. (3) The neighbors of the best available options now (score 3, 5) are added; penalties are assigned to the unselected moves. (4) Move with score 9 is selected after this sequence. If the synthesis is not ended here, this pathway is kept for further expanded by algorithm to find additional and possibly better solutions. (Figure 5)

6 Conclusion

A development of search strategy based on evaluation of reactivity on given reaction database, is focused in this work for the efficient generation of possible synthetic pathways towards complicated synthetic targets, and a global increasing of the likelihood of prediction success.

7 Reference

[1] "Planning chemical syntheses with deep neural networks and symbolic AI." Marwin H. S. Segler, Mike Preuss,& Mark P. Waller. Nature, 2018, 555, 604

[2] "Machine learning for organic synthesis: are robots replacing chemists?" Boris Maryasin, Philipp Marquetand,& Nuno Maulide. Angew. Chem. Int. Ed. (Highlight), 2018, 57, 6978

[3] "Prediction of organic reaction outcomes using machine learning." Connor W. Coley, et. al. ACS Cent. Sci., 2017, 3, 434

[4] "Computational chemical synthesis analysis and pathway design." Fan Feng, Luhua Lai, & Jianfeng Pei. Front. Chem. (Review), 2018, 6, 199

[5] "On the synthesis of machine learning and automated reasoning for an artificial synthetic organic chemist." Maneesh K. Yadav. New J. Chem. (Prospective), 2017, 41, 1411

[6] "Machine learning in computer-aided synthesis planning." Connor W. Coley, William H. Green, & Klavs F. Jensen. Acc. Chem. Res., 2018, 51, 1281

[7] "Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory." Tomasz Klucznik. et. al. Chem., 2018, 4, 522

[8] "Controlling an organic synthesis robot with machine learning to search for new reactivity." Jaroslaw M. Granda, et. al. Nature, 2018, 559, 377

[9] "Chemical synthesis with artificial intelligence: researchers develop new computer method." Phys. Org. (News), 2018, Mar 30.

[10] “Need to make a molecule? Ask this AI for instructions.” Nat. Res. (News), 2018, Mar 28.