

การพัฒนาโปรแกรมสำหรับออกแบบแผนการสังเคราะห์สารประกอบอินทรีย์โดยใช้ปัญญาประดิษฐ์

นายธวัชชัย จิตรพร

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาเคมี ภาควิชาเคมี
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2564
ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

PROGRAM DEVELOPMENT FOR SYNTHESIS PLAN DESIGN OF ORGANIC COMPOUNDS BY
USING ARTIFICIAL INTELLIGENCE

Mr. Tawatchai Jitporn

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Chemistry

Department of Chemistry

Faculty of Science

Chulalongkorn University

Academic Year 2021

Copyright of Chulalongkorn University

Thesis Title PROGRAM DEVELOPMENT FOR SYNTHESIS PLAN DESIGN OF OR-
GANIC COMPOUNDS BY USING ARTIFICIAL INTELLIGENCE
By Mr. Tawatchai Jitporn
Field of Study Chemistry
Thesis Advisor Associate Professor Somsak Pianwanit, Ph.D.

Accepted by the Faculty of Science, Chulalongkorn University in Partial Fulfillment
of the Requirements for the Master Degree

.....Dean of the Faculty of Science
(Professor Polkit Sangvanich, Ph.D.)

THESIS COMMITTEE

.....Chairman
(Professor Vudhichai Parasuk, Ph.D.)

.....Thesis Advisor
(Associate Professor Somsak Pianwanit, Ph.D.)

.....Examiner
(Associate Professor Kanet Wongrawee, Ph.D.)

.....External Examiner
(Assistant Professor Kiattisak Lugsanangarm, Ph.D.)

ธวัชชัย จิตรพร : การพัฒนาโปรแกรมสำหรับออกแบบแผนการสังเคราะห์สารประกอบอินทรีย์โดยใช้ปัญญาประดิษฐ์. (PROGRAM DEVELOPMENT FOR SYNTHESIS PLAN DESIGN OF ORGANIC COMPOUNDS BY USING ARTIFICIAL INTELLIGENCE) อ.ที่ปรึกษาวิทยานิพนธ์หลัก : รศ. ดร. สมศักดิ์ เพ็ญรวณิช, 25 หน้า.

การวางแผนสังเคราะห์โดยใช้คอมพิวเตอร์ช่วย (CASP) เป็นเครื่องมือทางคอมพิวเตอร์ที่ใช้ช่วยนักเคมีอินทรีย์ในการออกแบบและสังเคราะห์สารประกอบอินทรีย์โดยทำการคำนวณเส้นทางการสังเคราะห์ที่เป็นไปได้มากที่สุด เครื่องมือ CASP ส่วนใหญ่จะเป็นโปรแกรมเชิงพาณิชย์หรือโปรแกรมแบบไม่เปิดเผยต้นฉบับ ทำให้ไม่สามารถปรับแก้ไขเพื่อเพิ่มประสิทธิภาพหรืออัปเดตฐานข้อมูลได้ ดังนั้นจุดมุ่งหมายของงานนี้จึงเป็นการพัฒนาเครื่องมือ CASP แบบเปิดเผยต้นฉบับ ได้นำเอาคลังของหลักการสังเคราะห์ที่สร้างขึ้นจากฐานข้อมูลสิทธิบัตรยูเอสมาใช้และได้ทำการจัดหมวดหมู่หลักการใหม่เพื่อปรับปรุงความสามารถในการนำไปใช้ประโยชน์ จากนั้นนำไปใช้ในการฝึกฝนโครงข่ายหน่วยประสาทเทียม และได้โมเดลที่สำคัญ 2 โมเดลสำหรับโปรแกรมค้นหา ทำการสร้างเครื่องมือ CASP โดยใช้โปรแกรมค้นหานี้ร่วมกับวิธีค้นหาแบบแผนผังรูปต้นไม้มอนติคาร์โล เครื่องมือสามารถให้แผนการสังเคราะห์ของโมเลกุลเป้าหมายได้ 629 โมเลกุลจากทั้งหมด 1,237 โมเลกุล การปรับปรุงเพิ่มประสิทธิภาพของเครื่องมือน่าจะสามารถทำได้โดยการเพิ่มชุดข้อมูลปฏิกิริยาที่มีความเหมาะสมเข้าไปในคลังของหลักการสังเคราะห์

ภาควิชาเคมี..... ลายมือชื่อนิสิต.....
 สาขาวิชาเคมี..... ลายมือชื่อ อ.ที่ปรึกษาหลัก.....
 ปี การ
 ศึกษา2564.....

6072060123 : MAJOR CHEMISTRY

KEYWORDS : Computer-Assisted Synthesis Planning, Artificial Neural Network, Monte-Carlo Tree Search

TAWATCHAI JITPORN : PROGRAM DEVELOPMENT FOR SYNTHESIS PLAN DESIGN OF ORGANIC COMPOUNDS BY USING ARTIFICIAL INTELLIGENCE. ADVISOR : ASSOC. PROF. Dr. SOMSAK PIANWANIT, 25 pp.

Computer-Assisted Synthesis Planning (CASP) is a computational tool for facilitating organic chemists to design and synthesis organic compounds by calculating the best possible synthesis pathways. Most CASP tools are commercial or closed-source software and thus, they cannot be modified to improve performance or update database. Therefore, it is our aim to develop an open-source CASP tool. Library of reaction rules that was created from the US patent database was taken from literature and was then reclassified to improve its applicability. This modified library was then used for training neural network and 2 important models were obtained for the search engine. Using this search engine together with the Monte Carlo tree search method, our CASP tool can generate synthesis plan for 629 molecules out of 1237 target molecules. The performance of our tool could probably be further improved by including more qualified reactions dataset into the library of reaction rules.

Department :Chemistry..... Student's Signature.....

Field of Study :Chemistry..... Advisor's Signature.....

Academic Year :2021.....

ACKNOWLEDGEMENTS

The thesis for master degree could not have been completed without the good instruction and encouragement.

I would like to thank my supervisor Associate Professor Dr. Somsak Pianwanit for giving useful guidance, conceptualizing the research and revision during the time of my research.

Also I would like to thank my thesis committee, Professor Dr. Vudhichai Parasuk, Associate Professor Dr. Kanet Wongrawee, Assistant Professor Dr. Kiattisak Lugsanangarm for giving several suggestions. In addition, I would like to thank Dr. Connor W. Coley from Massachusetts Institute of Technology (MIT) for suggestion about this work.

I would like to thank Department of Chemistry, Faculty of Science, Chulalongkorn University for financial support. I would like to express my gratitude to the Center of Excellence in Computational Chemistry (CECC) for providing high-performance computing services and facilities.

Finally, I would like to thank to my family and friends both online and offline for physical and mental supports. Without these people, my thesis could not be successful or possible.

CONTENTS

	Page
ABSTRACT IN THAI	iv
ABSTRACT IN ENGLISH	v
ACKNOWLEDGEMENTS	vi
CONTENTS	vii
LIST OF TABLES	ix
LIST OF FIGURES	x
CHAPTER I Introduction	1
1.1 Retrosynthetic analysis	1
1.2 Computer-assisted synthesis planing	3
1.2.1 Core components of computer-assisted synthesis planning	3
1.2.2 Development of computer-assisted synthesis planning	5
CHAPTER II Theoretical Background	7
2.1 Artificial neural network	7
2.1.1 Perceptron	7
2.1.2 Optimization of neural networks	8
2.2 Monte-Carlo Tree Search (MCTS)	8
CHAPTER III Experimental Procedure	10
3.1 Component of CASP in this work	10
3.1.1 Library of reaction rules	11
3.1.2 Search engine	11
3.1.3 Database of starting material	11
3.1.4 Search strategy	11
3.1.5 Scoring function	12
3.2 Training neural network	12
3.2.1 Expansion model	12

	Page
3.2.2 In-scope filter model	13
3.3 Testing the CASP	13
CHAPTER IV Results and Discussion	15
4.1 Reaction template extraction	15
4.2 Neural network training	15
4.2.1 Expansion model	17
4.2.2 In-scope filter model	17
4.3 Synthesis plan performance	19
CHAPTER V Conclusion	22
VITAE	25

LIST OF TABLES

Table		Page
4.1	The top ten reaction templates extracted by RDChiral	16
4.2	The results at various rule frequency	18

LIST OF FIGURES

Figure		Page
1.1	Synthon and synthetic equivalent	2
1.2	Example of retrosynthetic analysis	2
1.3	CASP tool in SciFinder-n	3
1.4	Flowchart of CASP	5
2.1	Monte-Carlo Tree Search	9
3.1	Flowchart of experimental procedure	10
3.2	Example of target molecules	14
4.1	Training result of expansion model	17
4.2	Training results of in-scope filter model	19
4.3	The number of synthesis step that can be generated by this work	20
4.4	Time usage of each target molecule during generate synthesis plan	20
4.5	Example of synthesis plan that generated by this work	21

CHAPTER I

INTRODUCTION

Organic synthesis is one of the most important fields in organic chemistry. It focuses on finding synthesis pathway of organic compound or discovery novel reaction. To synthesize a new compound or a known compound, organic chemist must have solid knowledge and experience in organic synthesis⁰ to make synthesis plan reasonable. In most cases, organic chemist may have to do literature search to find more information about possible synthesis route. It is thus a difficult task for beginner in this field. This obstacle can be overcome logically by using retrosynthetic analysis in combination with computer-assisted synthesis planning tools.

1.1 Retrosynthetic analysis

Retrosynthetic analysis is a problem solving technique to generate synthesis plan in a reverse order. This process begins with analyzing a target molecule to find possible disconnectable functional group or bond. Applying this disconnection, the target molecule is separated into two smaller moieties, which are called “synthons” as shown in Figure 1.1. The structures of both synthons do not exist in reality so it is necessary to find compounds that can be used to generate these synthons and these compounds are called “synthetic equivalents”. Finding and applying disconnection are carried out recursively until synthetic equivalences become known purchasable compounds. The retrosynthetic analysis was initially logicalized in 1960s by E. J. Corey [1]. And since then, most of total synthesis research have been using this approach to generate synthesis plan and synthesize many compounds.

The example of retrosynthetic analysis can be shown in Figure 1.2. Although retrosynthesis analysis is very useful approach to find synthesis plan, it still requires strong background in organic synthesis to find suitable disconnection in each step. Therefore, computer technology comes to play a role to solve this drawback.

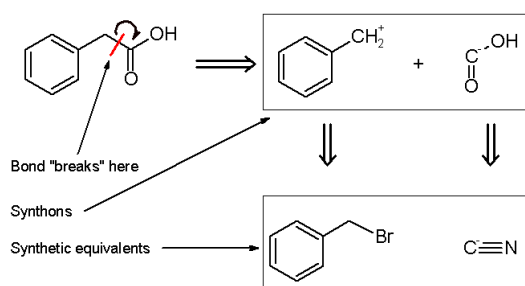


Figure 1.1: Synthon and synthetic equivalent

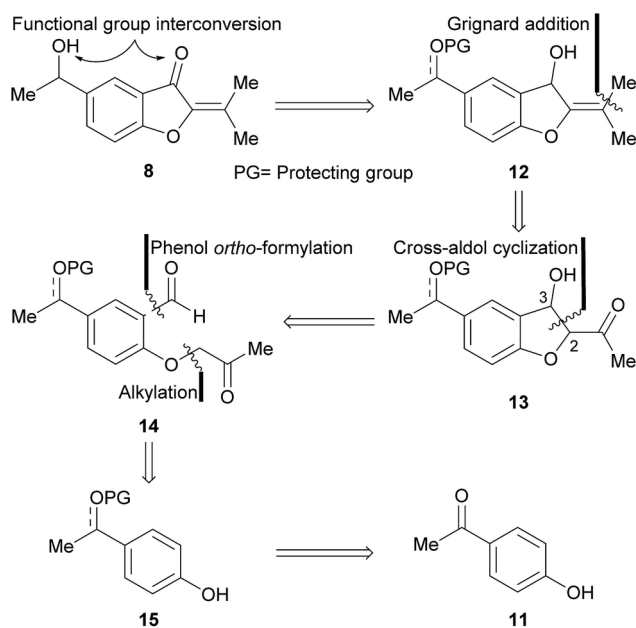


Figure 1.2: Example of retrosynthetic analysis

1.2 Computer-assisted synthesis planning

Computer-assisted synthesis planning (CASP) is a computational tool to generate synthesis pathway from target molecule. By using computer to calculate and find synthesis plan, it could greatly reduce time usage compared to generating synthesis plan manually. In addition, it is applicable for both novice and expert organic chemists. There are several available CASP tools, which were released as open-source or commercial software such as Reaxys and SciFinder-n which can be shown in Figure 1.3.

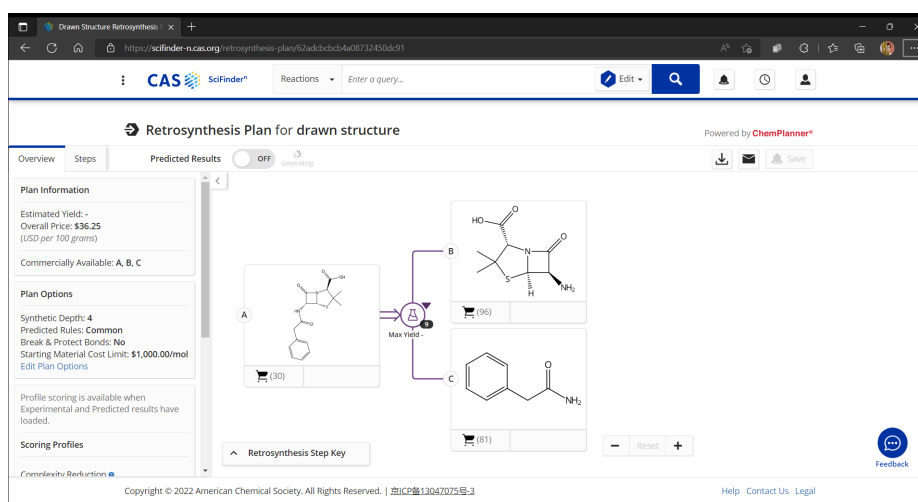


Figure 1.3: CASP tool in SciFinder-n

1.2.1 Core components of computer-assisted synthesis planning

Most of CASP tools consist of five components [2], which are explained as follows.

1. Library of reaction rules. The reaction rule is molecular fragment from both reactants and products that contains information about which atoms are disconnected or transformed from reactants into products or vice versa. The reaction rule can be encoded into SMARTS format or different format that component number 2 understand which part of molecule should be apply in this rule. The

library is used to serve as elementary synthesis step to build up the overall synthesis route of desired product. The library can be obtained by (a) using previous existing library and/or (b) creating a new one by mining the databases containing synthesis reactions.

2. Search engine. Starting from a target molecule as an input, search engine will determine disconnection by searching against library of reaction rules to find candidate reactions that can transform product (target molecule) backward one step to reactants. It will also construct reactant or product molecules according to selected reaction rule. Search engine is applied several times until the reactants are in the component number 3
3. Database of starting materials. This is normally a database of commercially purchasable compounds. If a reactant in any retrosynthesis step could be purchased, there is no need to step back to synthesize this compound. Therefore, this database is used to terminate retrosynthesis step cycle.
4. Searching strategy. Generally, synthesis route of most compounds composes of several synthesis steps from reactants to product. In each synthesis step, it is very common that there are several possible reactions with different reactants to synthesize a given product. In other word, several possible disconnections (several reaction rules) can be found by the search engine (component number 2) for a target product. So, there should be a good searching strategy to search through chemical space that can combine all synthesis steps together to give product. The retrosynthesis problem can be considered as a tree search problem, thus any tree search algorithm can be applied as searching strategy.
5. Scoring function. This function is used to calculate score of each possible synthesis route based on several factors, e.g., number of synthesis steps needed. User can use the score to evaluate which synthesis route is more feasible.

The workflow of CASP is shown in Figure **1.4**.

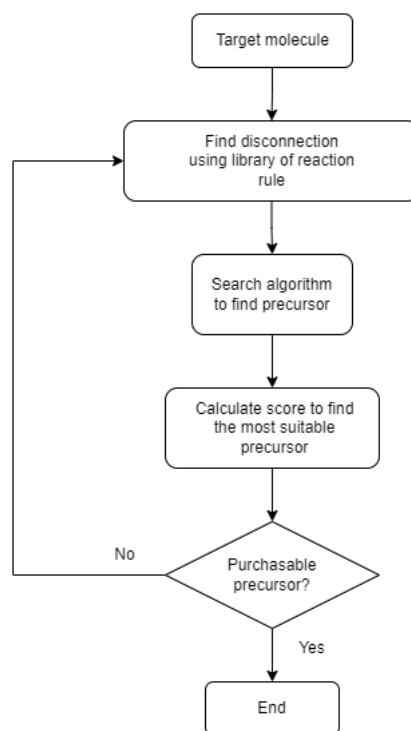


Figure 1.4: Flowchart of CASP

1.2.2 Development of computer-assisted synthesis planning

The development of CASP can be dated in 1960s from Dendral project [3] which aim to use artificial intelligence approach for generating synthesis plan but this task was unsuccessful. In 1967 E. J. Corey proposed the “retrosynthetic analysis” and systemized the rule [1]. This revolutionize the field to become more systematic and become the basis of CASP tools. Then Corey and Wipke presented the very first CASP tool called Organic Chemical Simulation of Synthesis but this project was short-lived and was split into two groups of CASP; LHASA for Logic and Heuristic Applied to Synthetic Analysis [4] and SECS for Simulation and Evaluation of Chemical Synthesis [5]. The LHASA project was developed by E. J. Corey, which used heuristic transforms written in chemical language called “CMTRN (Chemistry TRaNslator)”. However, one major drawback of this tool is a failure in dealing with reactions containing stereochemistry [6]. There were several CASP tools after LHASA but almost all of them were unsuccessful to use due to several problems which are discussed in details below.

1. Molecular context

Most of CASP tools use reaction template coded by expert and this can be problematic when molecular context coming to play. Even similar molecules but with different functional groups can lead to different results. This gives a similar role about regioselectivity problem.

2. Size of chemical space and searching algorithm

The chemical space for synthesis planning is not much large but still a challenging due to most of CASP tools use exhaustive search or simple best-first search which can be a problem when searching an enormous space.

3. Synthetic position

Unlike chess or any other games, the synthesis position is ill-defined problem which means it cannot be defined easily which position should be applied or which reaction should be used.

As the review paper was published in 2016 [7], the CASP is still the challenging problem due to several points that were already discussed. In 2017, the preprint of developing of CASP tool was released and this mark as significant because this can solve most of the previous problems. And the next year this preprint article was published into research article [8]. In that research article, the data-driven approach was used to extract data in synthesis database into reaction template and then it was combined with three models of artificial neural network and Monte-Carlo tree search algorithm (MCTS). After 2018, CASP development have been using this approach by employing artificial neural network and Monte-Carlo tree search (MCTS) algorithm with their own dataset for training neural network [9].

CHAPTER II

THEORETICAL BACKGROUND

2.1 Artificial neural network

Artificial Neural Network or ANN is the mathematical model that has inspiration from biological nerve cell. Initialized in 1943 by Warren McCulloch and Walter Pitts [10]. Artificial neural network model can be used as single unit called perceptron or many units connected to each other. The many units of perceptron that connect to as a layers sometime called “deep learning”. Artificial neural network model can be applied in many field such as object detection, image processing, natural language processing etc.

2.1.1 Perceptron

The perceptron is the smallest unit of artificial neural network. Introduced by Frank Rosenblatt in 1958 [11]. The equation of perceptron is shown in equation 2.1.

$$y_i = f(w_1x_1 + w_2x_2 + w_3x_3 + \dots + w_nx_n + b) \quad (2.1)$$

where x_i are the input of perceptron, w_i are weight or parameter of perceptron, b is bias, and $f(x)$ is activation function. The perceptron can solve logical problem such as “and” gate and “or” gate. However it cannot solved the exclusive or gate problem with single perceptron. This problem lead to the use of multilayer perceptron, which means using another perceptron connect with previous perceptron and stack into layer, hence the name multilayer perceptron.

2.1.2 Optimization of neural networks

To make neural network learning from data that feed into model, the parameter of neural network must be optimized. In this case, the optimization algorithm can be used to optimize the parameters which are weights and biases in neural network model. The very first algorithm to optimize parameter of neural network is gradient descent method which is shown in equation 2.2.

$$w_{i+1} = w_i - \alpha \frac{dJ}{dw} \quad (2.2)$$

while w_i are the parameter in previous step, α is learning rate, and $\frac{dJ}{dw}$ are the derivative of cost function by parameters. The problem is how to calculate the derivative of these functions since there are several parameters to calculate and it is difficult to calculate.

Yann LeCun proposed the back propagation algorithm [12]. The back propagation algorithm uses a chain rule in calculus for calculating the weights and biases by back-propagating the derivative of previous function continue to weights and biases. This is shown in equation 2.3.

$$\frac{dJ}{dw} = \frac{dJ}{df} \frac{df}{dz} \frac{dz}{dw} \quad (2.3)$$

However, the gradient descent method is no longer used for entire dataset because it is computational expensive. Thus, stochastic gradient descent method and variation will be use instead such as stochastic with momentum, RMSProp, and Adam.

2.2 Monte-Carlo Tree Search (MCTS)

Monte-Carlo Tree Search or MCTS is the tree search algorithm, purposed by Rémi Coulom in 2006 [13]. Monte-Carlo tree search algorithm differ than traditional tree search by using random simulation (Monte-Carlo simulation) as part of tree search hence the name.

This algorithm consists of four phases.

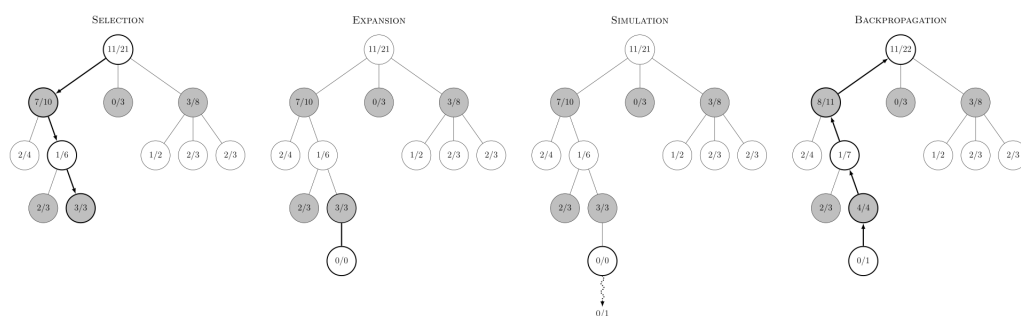


Figure 2.1: Monte-Carlo Tree Search

1. Selection

This phase will select node from the root node until the leaf node is reached

2. Expansion

This phase will create new node to expand the leaf node from previous phase

3. Rollout or Simulation

In this phase, a random node will be build and doing Monte-Carlo simulation until the target node is reached.

4. Update

This final phase will update the reward or score from the simulation result and send it back to parent node until root node is reached.

The advantage of Monte-Carlo tree search is the tree can grow asymmetrically which is useful for searching on tree problem with high branching factor. However, the drawback of Monte-Carlo tree search is it can lead into loss state due to policy of selective node expansion. Monte-Carlo tree search can be applied in game such Go [14, 15] and retrosynthesis problem [8].

CHAPTER III

EXPERIMENTAL PROCEDURE

3.1 Component of CASP in this work

In this thesis work, the template-based data-driven approach was employed. All the components from section 1.2.1 are described in details on how to get these components. The flow of our work is schematically shown in Figure 3.1.

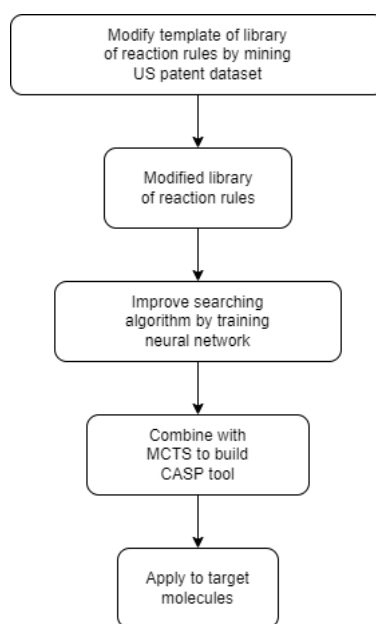


Figure 3.1: Flowchart of experimental procedure

3.1.1 Library of reaction rules

This work used the library of reaction rules developed by D. Löwe in his Ph.D. dissertation [16] as a template. The library was extracted from the US patent data [17]. There are two databases: granted patents from 1976 to September 2016 and applications from 2011 to September 2016. These two databases were combined into one dataset and duplicate reactions were drop out. Next step is using modified version of RDChiral [18] reaction-extraction function in python library to extract reaction center, reactant and product molecules. Reaction that cannot be extracted and reaction with unique template will be discarded. These reactant and target molecules were used to create library of reaction rules and will be used to train neural network afterward.

3.1.2 Search engine

This component used RDKit [19] to generate reactant from product that corresponds with reaction center. With the help of neural network to guide what reaction center should be used and confirmation of reaction by using another neural network model to get reasonable reactant molecule.

3.1.3 Database of starting material

This database was created by combining ZINC dataset [20], which is a free database of commercially available compounds, and the two sources of starting material suppliers which are AlfaAesar and Acros.

3.1.4 Search strategy

Searching strategy in this work used Monte-Carlo Tree Search, which is similar to previous work [8]. However, instead of using three models of neural network, only

expansion model and in-scope filter neural network model were used in this work. The reason to omit rollout model is that it has similar role to expansion model, which is to calculate what probability of each reaction should be used but in different phase of MCTS so to save training time only expansion model was used in this work.

3.1.5 Scoring function

The scoring function in this work was modified from Segler's work [8]. It is shown in below.

$$Q(s, a) = \frac{1}{N(s, a)} \sum_{i=1}^n z_i W(b_i) \quad (3.1a)$$

$$W(b_i) = \max \left(0, \frac{L_{max} - \xi(b_i)}{L_{max}} \right) \quad (3.1b)$$

$$\xi(b_i) = \text{length}(b_i) - \sum_{j=1}^J k P(s_j, a_j) \quad (3.1c)$$

while $N(s, a)$ is number of times that explored, L_{max} is max length of synthesis step which was set to 10, $P(s, a)$ is prior probability that is calculated by neural network, k is damping factor which was set to 0.99, z_i is the reward of state.

3.2 Training neural network

3.2.1 Expansion model

Expansion model is neural network model for using in expansion phase of Monte-Carlo tree search to find which reaction rule should be applied. This model is similar to previous work [8].

To get the training data for this model, all extracted reaction rules were used. The

product of each reaction was labeled according to what reaction rule came from since several reactions can give the same reaction rule. The number of product from reaction per reaction rule was varied from 5 to 100 with increment by 5, since this model is important for classifying on what reaction should be use so this model need to be varied to make sure that the optimized condition for accuracy and number of frequency of reaction rule can be found.

3.2.2 In-scope filter model

In-scope filter model is neural network model for using in expansion phase to validate molecule after applied search engine that corresponding to reaction rule. This model is similar to previous work [8].

To get the training data for this model, all reactions in the dataset that were labeled from previous step and reaction rules were used. All reactant were applied by each reaction rule to generate product, if product generated from reaction rule is the same molecule as labeled product, this reaction will be labeled as “corrected” reaction, otherwise it will labeled as “incorrected” reaction. This will generate a lot of dataset and also unbalance due to the number of correct reaction is much less than number of incorrect reaction. To maximize the accuracy of this model, the incorrect reaction will be sampling according to number of correct reaction.

3.3 Testing the CASP

To test our CASP method, the dataset from real target molecules must be chosen. Target molecules to test this CASP are target molecules taken from publication in Journal of Medicinal Chemistry in 2019.

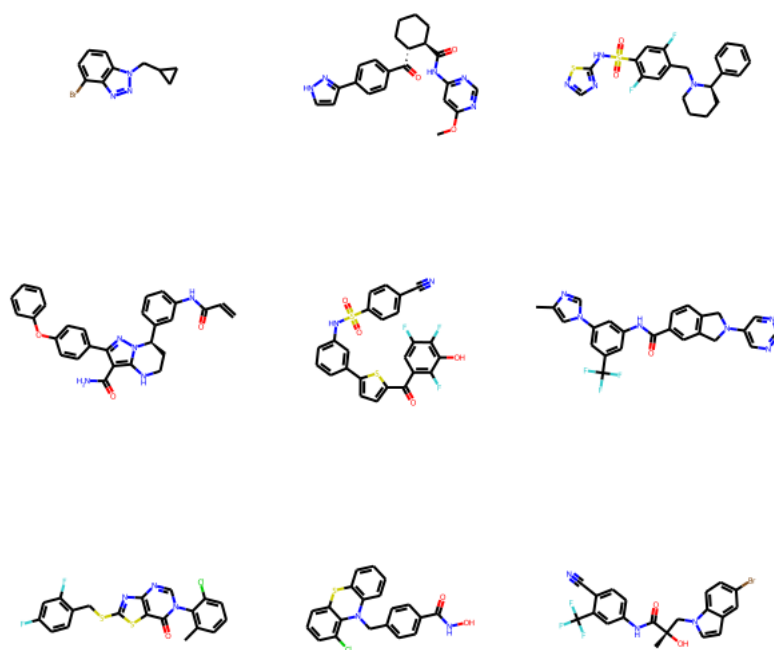


Figure 3.2: Example of target molecules

CHAPTER IV

RESULTS AND DISCUSSION

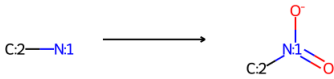
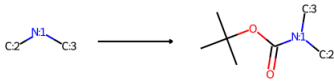
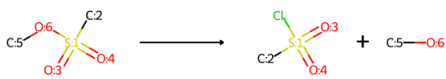
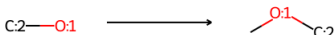
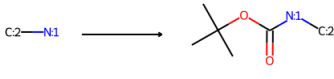
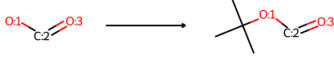
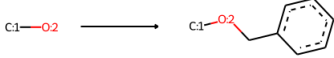
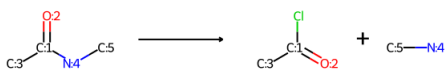
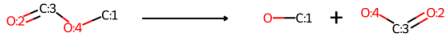
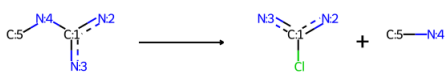
4.1 Reaction template extraction

From 1,484,441 reactions in the US Patent dataset, only 644,619 reactions extracted which were 43.42 percent. The major reason that several reactions cannot be extracted is some reactions have incorrect SMILES format when extracted from literature [17]. Some reactions have atom labels but they do not changed in both reactant and product and thus a reaction center could not be identified. All the extracted reactions were classified into groups, which are called reaction templates. Totally 1,225 reaction templates were obtained. Top ten reaction templates with the highest frequency are shown in Table 4.1. It is worth to mention that several reactions gave unique reaction template which cannot be used to train in neural network model due to insufficient data per template. Within 1,225 reaction templates, there are only 29 reaction templates that contain stereochemistry which is just 2.37 percent. This indicated that there are not much stereochemistry in reaction template. The reason is probably because there is not much stereochemistry in reaction dataset with perfect reaction dataset, also the stereochemistry is not in the reaction center which is omitted by the RDChiral itself.

4.2 Neural network training

Neural network training results are discussed here.

Table 4.1: The top ten reaction templates extracted by RDChiral

reaction template	number of reaction
	11,470
	6,507
	3,764
	3,745
	3,587
	2,825
	2,148
	1,932
	1,870
	1,869

4.2.1 Expansion model

The training results for expansion model are shown in Figure 4.1. The optimal condition of this model is at epoch one from epoch zero which gives the lowest cost function value of validation data which is 1.4228. The accuracy of this model is 63.59 percent which is similar to the previous work [8]. Since this model is multiple classification, precision and recall of this model were not calculated due to numerous number of class and its difficulty in interpretation. Also, the number of optimal reaction per class at 50 gave the optimal condition satisfied with number of classes, number of reaction per class, accuracy of model and performance of this CASP which are shown in Table 4.2. This number was also used in previous work in the rollout model [8].

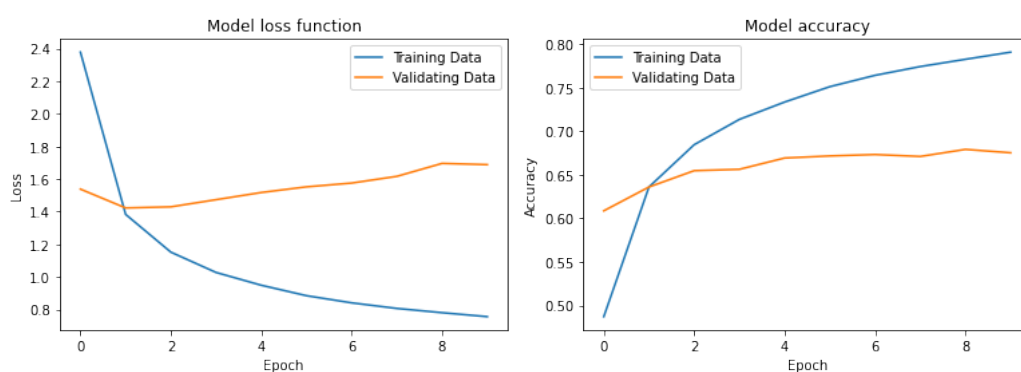


Figure 4.1: Training result of expansion model. Left is loss function of expansion model, and right is accuracy of this model.

4.2.2 In-scope filter model

The training results of in-scope filter model are shown in Figure 4.2. The optimal condition of this model is at epoch six from epoch zero which gives the lowest cost function value of validation data which is equal to 0.3740. This model is binary classification model, the precision and recall of this model were calculated. The precision is 0.8373 and the recall is 0.8525.

Table 4.2: The results at various rule frequency

Rule frequency	Number of rule	Model Accuracy (%)	CASP Performance (%)
5	14,720	52.27	40.18
10	7,003	56.52	42.28
15	4,528	60.17	43.41
20	3,337	61.51	46.24
25	2,626	61.73	47.53
30	2,125	62.18	51.58
35	1,813	61.55	48.99
40	1,535	63.04	49.80
45	1,362	63.67	52.14
50	1,225	65.02	50.44
55	1,094	63.80	46.40
60	992	65.56	51.58
65	911	65.31	53.03
70	842	66.91	51.41
75	783	67.43	49.88
80	728	67.81	49.07
85	670	68.42	45.59
90	626	66.07	51.58
95	581	67.91	51.41
100	553	68.20	51.66

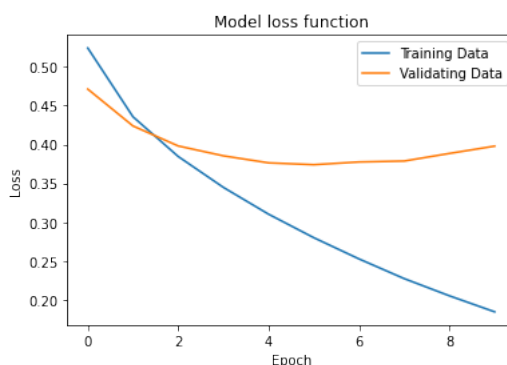


Figure 4.2: Training results of in-scope filter model

4.3 Synthesis plan performance

Using 1,237 target molecules in the testing test, our CASP tool can successfully generate synthesis plan for 629 target molecule, which is 50.85 percent. According to Figure 4.3 most of successful synthesis plans consist of 1 - 2 step(s) but several target molecules can reach up to the maximum number of 10 steps. According to Figure 4.4, the time usage for this CASP is just 1 - 2 second(s) although there are several target molecules that need 20 seconds. The reason is the calculations of these molecules reached the maximum number of reaction step which is set to 10. The example of synthesis plan that was generated by this work is shown in Figure 4.5. Although there are 1,225 reaction templates, this number is still not enough for generate high quality synthesis plan because some of target molecules might contain some functional groups that are actually transformable but they are not included in this work.

In order to compare the performance of our CASP tool with the commercial one, 100 target molecules were randomly selected from our testing set and were inputted into the SciFinder-n tool. The SciFinder-n could generate synthesis plan for 76 molecules. It could be possible to further improve the performance of our tool by including more qualified reactions dataset into the library of reaction rules.

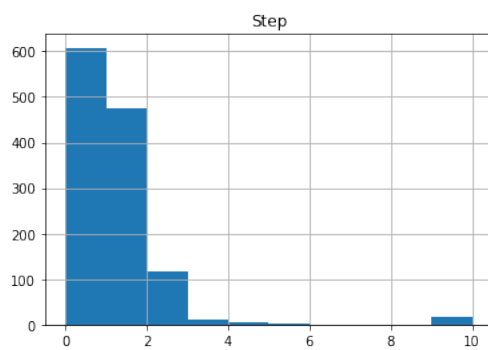


Figure 4.3: The number of synthesis step that can be generated by this work

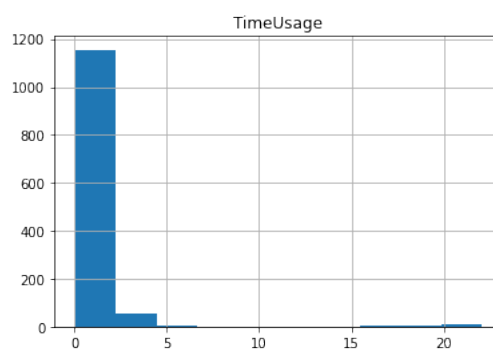


Figure 4.4: Time usage of each target molecule during generate synthesis plan

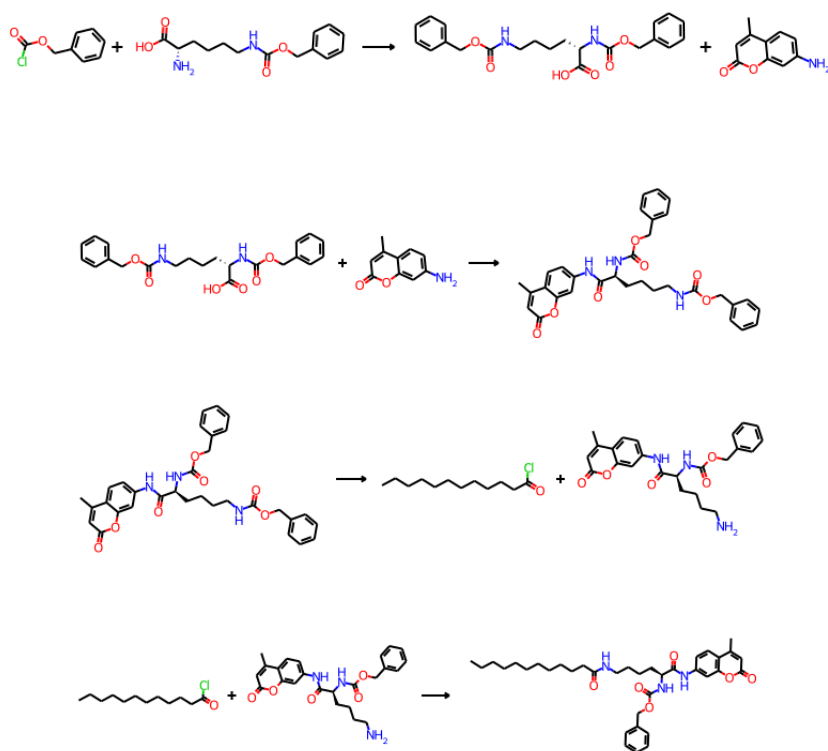


Figure 4.5: Example of synthesis plan that is generated by this work. This target molecule can be synthesized within 4 steps.

CHAPTER V

CONCLUSION

The open-source CASP tool was successfully developed by using neural network in combination with MCTS. The accuracy of expansion model is 63 percent by using reaction per class at 50, precision is 0.84 and recall is 0.85. The number of reaction center from this work is 1,225. The tool can generate synthesis plan of 629 from 1237 target molecules. All the source code are available at https://github.com/tjthecalculator/thesis_retrosynthesis

Suggestion for further work is the improvement of performance by including more high-quality reactions dataset into the library of reaction rules. In addition, it would be very convenient for users if other formats for structure of input, apart from SMILES format, are allowed. The improvement could also be done by applying neural network to find disconnection. This need to be investigated and studied.

REFERENCES

- [1] Corey, E.J. General methods for the construction of complex molecules. Pure and Applied Chemistry 14 (1967): 19–37.
- [2] Coley, C.W., Green, W.H. and Jensen, K.F. Machine Learning in Computer-Aided Synthesis Planning. Accounts of Chemical Research 51 (2018): 1281–1289.
- [3] Lindsay, R.K., Buchanan, B.G., Feigenbaum, E.A. and Lederberg, J. DENDRAL: A case study of the first expert system for scientific hypothesis formation. Artificial Intelligence 61 (1993): 209–261.
- [4] Corey, E.J., Wipke, T., Cramer III, R.D. and Howe, J. Computer-assisted synthetic analysis. Facile man-machine communication of chemical structure by interactive computer graphics. Journal of American Chemical Society 94 (1972): 421–430.
- [5] Wipke, T., Ouchi, G.I. and Krishnan, S. Simulation and evaluation of chemical synthesis—SECS: An application of artificial intelligence techniques. Artificial Intelligence 11 (1978): 173–193.
- [6] Hendrickson, J.B. and Toczko, G. SYNGEN program for synthesis design: basic computing techniques. Journal of Chemical Information and Computer Sciences 29 (1989): 137–145.
- [7] Szymkuc, S., Gajewska, E.P., Klucznik, T., Molga, K., Dittwald, P., Startek, M., Bajczyk, M. and Grzybowski, B.A. Computer-Assisted Synthetic Planning: The End of the Beginning. Angewandte Chemie International Edition 55 (2016): 5904–5937.
- [8] Segler, M., Preuss, M. and Waller, M.P. Planning Chemical Synthesis with Deep Neural Network and Symbolic AI. Nature 555 (2018): 604–610.
- [9] Genheden, S., Thakkar, A., Chadimova, V., Reymond, J.L., Engkvist, O. and Bjerrum, E. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. Journal of Cheminformatics 12 (2020): 1–9.
- [10] McCulloch, W.S. and Pitts, W. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics 5 (1943): 115–133.
- [11] Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review 65 (1958): 386–408.

- [12] LeCun, Y. a Learning Scheme for Asymmetric Threshold Networks. In Proceedings of Cognitiva 85 (1985).
- [13] Coulom, R. Efficient selectivity and backup operators in Monte-Carlo tree search. In Proceedings Computers and Games 2006. Springer-Verlag (2006).
- [14] Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T. and Hassabis, D. Mastering the game of Go with deep neural networks and tree search. Nature 529 (2016): 484–489.
- [15] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van der Driessche, G., Graepel, T. and Hassabis, D. Mastering the game of Go without human knowledge. Nature 550 (2017): 354–359.
- [16] Lowe, D. Extraction of chemical structures and reactions from the literature (Doctoral thesis). Ph.D. thesis, University of Cambridge (2012).
- [17] Lowe, D. Chemical reactions from US patents (1976-Sep2016). [Online]. 2017. Available from: <https://doi.org/10.6084/m9.figshare.5104873.v1>
- [18] Coley, C.W., Green, W.H. and Jensen, K.F. RDChiral: An RDKit Wrapper for Handling Stereochemistry in Retrosynthetic Template Extraction and Application. Journal of Chemical Information and Modeling 59 (2019): 2529–2537.
- [19] Landrum, G. RDKit: Open-source cheminformatics. [Online]. 2010. Available from: <https://www.rdkit.org>
- [20] Sterling, T. and Irwin, J.J. ZINC 15 – Ligand Discovery for Everyone. Journal of Chemical Information and Modeling 55 (2015): 2324–2337.

VITAE

Personal Details

Name	Mr. Tawatchai Jitporn
Date of Birth	10/10/1994
Place of Birth	Bangkok, Thailand
Address	Bangkok, Thailand
E-mail address	mr.tawatchai@live.com

Education

2017-2022	M.Sc. in Chemistry, Chulalongkorn University, Thailand
2013-2017	B.Sc. in Chemistry, Chulalongkorn University, Thailand