University of Southern Denmark

Department of Mathematics
and Computer Science

Master Thesis – 31/12-2017

# Computational Synthesis Planning Using Big Data

Computational Syntese planlægning ved hjælp af big data

*Author:*
Henrik Schulz

*Supervisors:*
Daniel Merkle

## Abstract

## Resumé

## Contents

## 1 Acknowledgements

## 2 Introduction

kkk [2]

### 2.1 Overview

Hvad indeholder hver sektion??

## 3 Preliminaries

This section contains definitions that will be used throughout this paper. It is assumed that the reader have a basic understanding of graph theory.

**Hypergraphs**
A directed hypergraph $h$ is a set of $V$ of vertices and a set $E$ of hyperedges, where each hyperedge $e = (T(e), H(e))$ is an ordered pair of non-empty multi-sets of vertices. The set $T(e)$ is denoted as the tail of the hyperedge and $H(e)$ is the head. If $|H(e)| = 1$ then the hyperedge is denoted as a B-hyperedge. If all edges in the hypergraph is B-hyperedges, then the graph is denoted a B-hypergraph. This paper will only consider hypergraphs that are B-hypergraphs. A hypergraph $H' = (V', E')$ is a subhypergraph of $H = (V, E)$ if $V' \subset V$ and $E' \subset E$.[3]

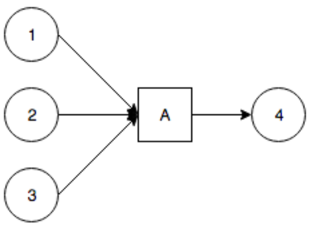


Fig. 1: Example of hyperedge $A$. $T(A) = \{1, 2, 3\}$, $H(A) = \{4\}$

**Hyperpaths**
A path $P_{st}$ from $s$ to $t$ in a B-hypergraph is a sequence $P_{st} = \langle e_1, e_2, e_3, ..., e_q \rangle$ of B-hyperedges such that $s \in T(e_1)$ and $t = H(e_q)$ and $H(e_i) \in T(e_{i+1})$ for $i = 1..q-1$. Its length $|P_{st}|$ is the number $q$ of hyperedges. If $t \in T(e_1)$, then $P_{st}$ is a cycle. A hypergraph is acyclic if it does not contain any cycles. [3]

## 4 Finding The K-Best Synthetic Plans

This section describes how to find the K-Best paths of a hypergraph modifying an algorithm made by Jin Y. Yen. The section starts out by describing the work flow of the algorithm, and then proceeds to deal with the alterations that are needed to run the algorithm on a hypergraph.

## 4.1 Yen's Algorithm

Yen's algorithm is an algorithm that computes the K-shortest paths for a graph with non-negative edges. It was publish in 1971 and uses any shortest path algorithm to find the best path and then proceeds to find the $K-1$ deviation of the best path. [4]
It starts out by finding the best path using a shortest path algorithm. Once the best path have been found it uses the path to find all the potential second best paths by fixing and removing edges in the graph.
By using the same first vertex as the original path but removing the first edge, it forces the shortest path algorithm to take another route through the graph and thereby creating a potential second best path. This is added to the list of potential paths and the algorithm can continue to derive other paths from the best path. By fixing the first edge in the previous best plan, Yen's algorithm forces the shortest path algorithm to take the first edge which it now shares with the best path. However, now the algorithm have removed the second edge from the original path and once again forces the shortest path algorithm to find alternative routes. This process is then repeated until we reach the next to last vertex in the best path.
By sorting the list of potential paths, it has the second best path at the start of the list and it can add it to the final list of best path. The algorithm then repeats on the second best path to find the third best path. This is done until all K-best path have been found or there are no more paths to find.

## 4.2 Yen's Algorithm On Hypergraphs

We use the principles from Yen's algorithm to make our own algorithm that will work on hypergraphs. To handle the problem of generating all derived paths from our best path in our hypergraph, we use a method called Backwards-Branching. [3] [5] [6]

Algorithm 1: Backwards Branching for B-Hypergraph

```
function Back−Branch(H,$\pi$)
    B=$\emptyset$
    for i = 1 to q do
        Let H$^i$ be a new hypergraph
        H$^i$.V = H.V
        // Remove hyperarc from H
        H$^i$.E = H.E \{$\pi.p(v_i)$}
        // Fix Back tree
        for j = i+1 to q do
            H$^i$.BS(vj) = \{$\pi.p(v_j)$\}
        B = B $\cup$ \{H$^i$\}
    return B
```

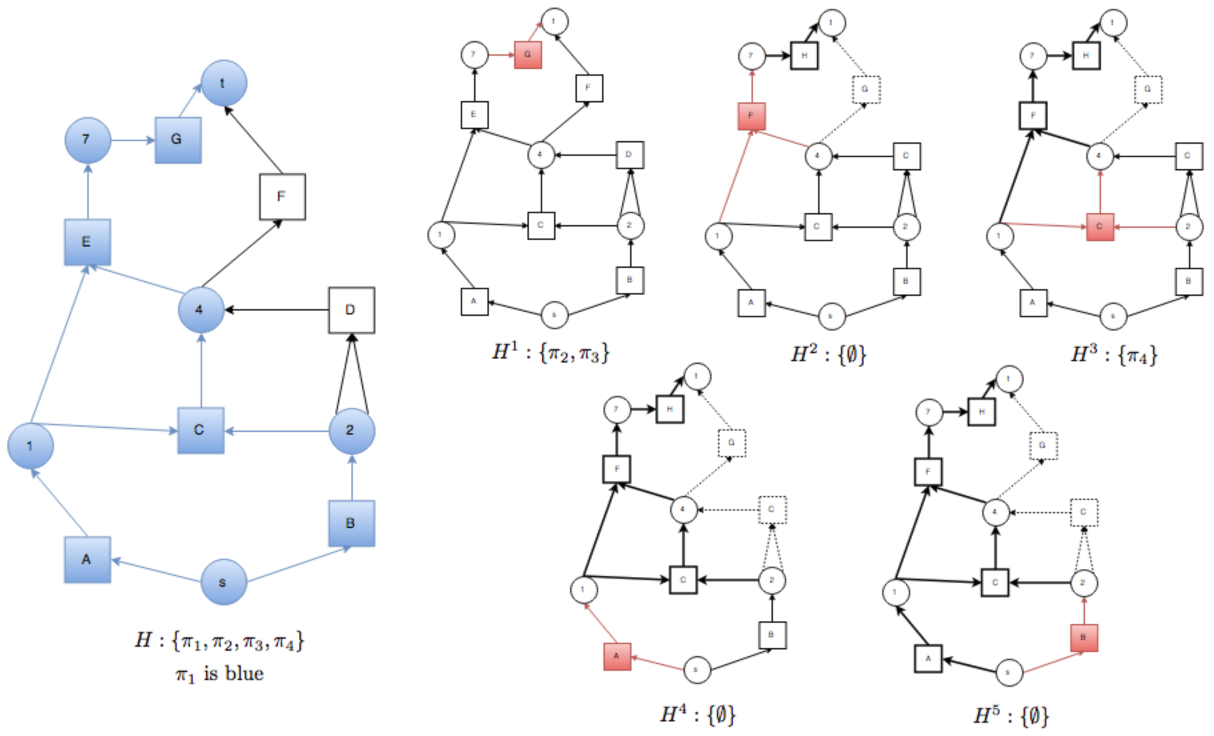


Fig. 2: An example of branching for hypergraphs. $H$ is the original hypergraph. The vertices and hyperedges marked with blue are part of the best path. The rest of the figure illustrates the backwards branching. Each of the smaller figures shows a hypergraph $H^i$ and how it is created from $H$. Dotted hyperedges and red heyperedges are not part of the hypergraph, but have been deleted due to branching. Hyperedges in bold are fixed hyperedges. When hyperedges are fixed it leads to other hyperedges being deleted (dotted). The caption beneath each hypergraph represents the possible paths that are available in the given hypergraph.

However, this algorithm have a problem when working on a larger hypergraph. It demands that each time we make alterations on the hypergraph we have to make a copy, $H^i$, of the graph, $H$, with the exception of the hyperedges that is removed when fixing the back tree and removing $\pi.p(v_i)$.
This could easily work for smaller graphs, but if we use this on the hypergraph that we generate from the beilstein database, we would have to copy a graph of multiple GigaBytes.
To handle this problem I came up with the idea of creating an overlay for the graph instead of copying it. The overlay would work as an transparent on top of the original graph, stating which edges still were accessible. This is done by creating a *vector<bool>* which has a length of $R$, where $R$ is the number of reactions. Normally a reaction would contain at least 24 bytes of data:

- 2x ints of 4 bytes each

- 1x double of 8 bytes
- 1x *vector<int>* head of length one of at least 4 bytes
- 1x *vector<int>* tail of length $N$ (number of educts) of at least 4 bytes

This can be reduced dramaticly by using the *vector<bool>*, since c++ only uses 1 bit per boolean in the vector instead of the regular 1 byte per boolean.[7] This means if working on a hypergraph with 40 million reactions, we would be able to create an overlay using 5 MB of space per alternated graph, instead of copying a hypergraph were the reactions alone uses at least 960 MB per copy. As the figure below shows, we never change or remove anything on the hypergraph. We simply create the following overlay:

| Reaction | A | B | C | D |
|---|---|---|---|---|
| Usable | true | true | true | false |
| Bit Representation | 1 | 1 | 1 | 0 |

And then when trying to use an edge, we ask: "Does overlay at reaction A exist?". If yes, you can use it. If no, the edge have been "removed", and therefore cannot be used.

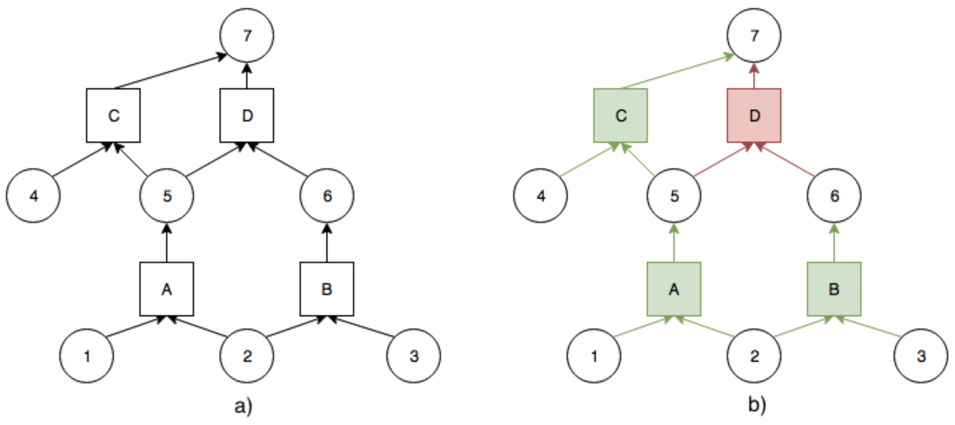


Fig. 3: a) The original hypergraph. b) The original hypergraph, but using the overlay. If the reaction is green it is still usable. If red then it have been "removed" from the hypergraph.

This of course means that the algorithm for back-branch have to be changed accordingly. Instead of the hypergraph as input we now give it an overlay. This overlay is changed so it fits with the new layout of the graph. Instead of deleting hyperedges in the copy, we now simply changes the boolean at the index of the hyperedge.id. True if we should add the hyperedge and false when we want to "remove" a hyperedge.

Algorithm 2: Backwards Branching for B-Hypergraph using overlay

```
function BackwardsBranching($\pi$, Overlay)
```

```
    List B = $\emptyset$
    // q = Path length
    for i = 1 to q do
        //remove i'th hyperedge from Path in overlay
        Set Overlay[$\pi$[i]] to false
        //fix the backtree
        for j = i downto 1  do
            vertex C <− $\pi$[j].head
            for each hyperedge into C
                Set Overlay[reaction.id] to false
            Set Overlay[$\pi$[j]] to true
        endfor
        B = B $\cup$ $\{Overlay\}$
    endfor
    return B
endfunction
```

Now that we have the branching in place are we able to construct an algorithm that are similar to Yen's algorithm but can run on hypergraphs. As input it takes a start vertex, a goal vertex, and an integer $K$, where $K$ is the number of best paths we want to find.
It creates a heap, $L$, and a list, $A$, which will contain the $K$-best paths once the algorithm is finished. It then finds the best path using a shortest path algorithm and inserts it into the heap. Inside the loop it extracts the best path found from the heap and performs a backward branching, and finds all possible paths in the branches. If there is a path from $s$ to $t$, then this path is added to the heap. The algorithm either terminates if the heaps is empty ( No more paths available) or once it have found the $K$-best paths.

Algorithm 3: K-Shortest Paths Algorithm in B-Hypergraph

```
function YenHyp(s, t, K)
    L = new heap with elements (overlay, $\pi$)
    A = List of shortest paths
    //(Graph is default overlay (all true))
    $\pi$ = shortestPath(Graph, s,t)
    Insert (Graph, $\pi$) into L
    for k = 1 to K do
        if L = $\emptyset$
            Break
        endif
        (Overlay$'$, $\pi'$) = L.pop
        add $\pi'$ to A
        for all Overlay$^i$ in BackwardBranching((Overlay$'$,$\pi'$)) do
            $\pi^i$ = shortestPath(Overlay$^i$, s, t)
            if $\pi^i$ is complete
                Insert( $H^i$, $\pi^i$) into L
            endif
        endfor
    endfor
    return A
```

```
endfunction
```

## 5 Shortest Path

This section describes two different approaches to the shortest path problem in a hypergraph. A dynamic approach proposed by Carsten Grønbjerg Lützen and Daniel Fentz Johansen [6] and a Dijkstra inspired approach proposed by Lars Relund Nielsen, Kim Allan Andersen and Daniele Pretolani [5].

## 5.1 Dynamic Approach

kkk [6]
kkkk [3]

### 5.1.1 Approach

Algorithm 4: Dynamic programming for finding the best path

```
function Min(V)
    if(V) is starting material
        return Cost of V
    endif
    mincost <− inf
    for all $e \in BS(V)$ do
        cost <− cost of e
        for all $u \in Tail(e)$ do
            cost <− cost + Min(u)
        endfor
        if mincost $\leq$ cost
            mincost <− cost
            V.minedge <− e
        endif
    endfor
    return mincost
endfunction
```

### 5.1.2 Testing

### 5.1.3 Problems

## 5.2 Nielsens Algorithm

k [5]

### 5.2.1 Approach

Algorithm 5: Dynamic programming for finding the best path

```
Initialization: Set W(u) = inf $\forall u \in V$, $k_j = 0 \forall e_j \in E$,$Q = \{s\}$ and $W(s) = 0$
function SBT−Dijkstra
    while ($Q = \emptyset$) do
        select and remove $u \in Q$ such that $W(u) = min\{W(x)|x \in Q\}$
        for ($e_j \in FS(u)$) do
            $k_j < -kj + 1$
            if ($k_j = |T(e_j)|$)
                v <− h($e_j$)
                if (W(v) > w($e_j$) + F($e_j$))
                    if ($v \notin Q$
                        Q <− Q ∪ \{v\}
                    endif
                W(v) <− w($e_j$)+ F($e_j$)
                p(v) <− $e_j$
                endif
            endif
        endfor
    endwhile
endfunction
```

#### 5.2.2 Testing

#### 5.2.3 Problems

#### 5.2.4 Optimizing

# 6 Work With Beilstein Data

This section HOFFMANN, REAXYS

## 6.1 The Database

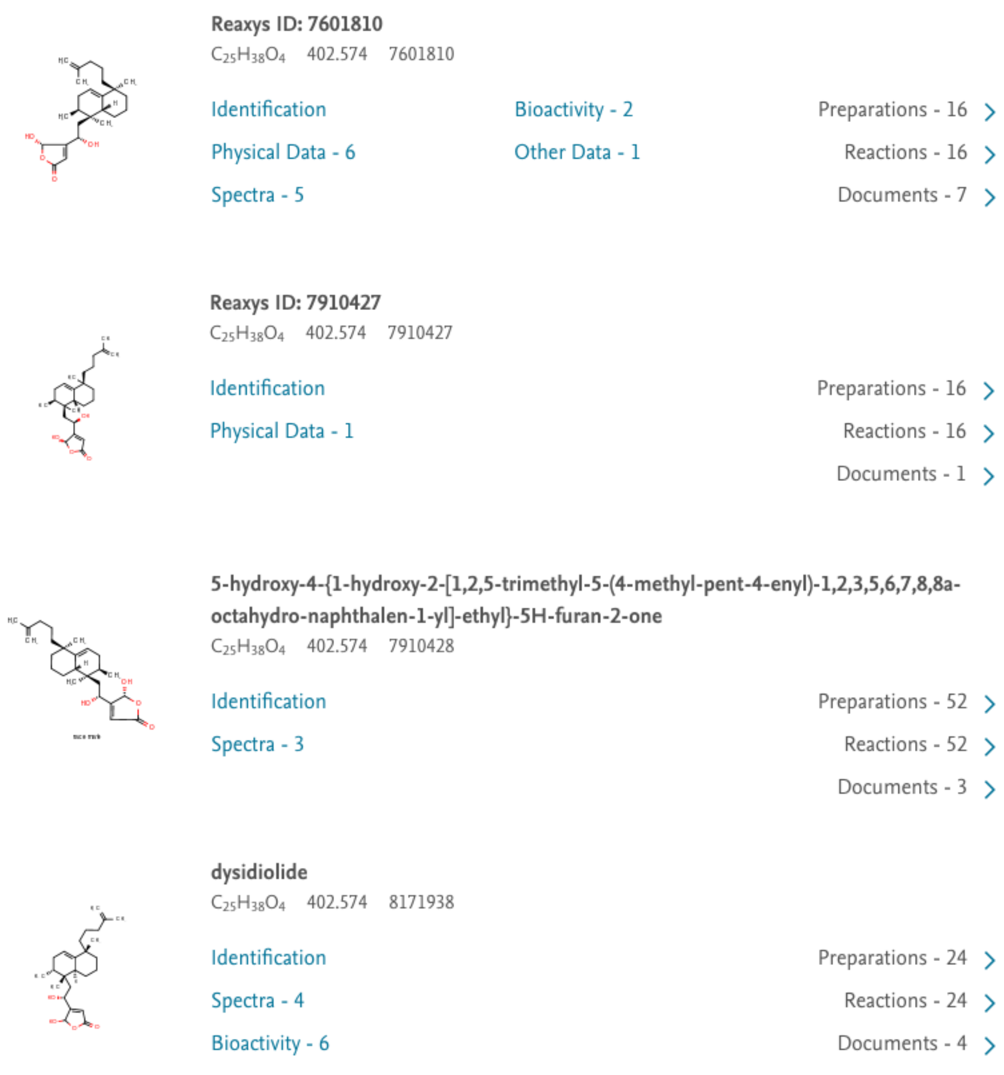
The Beilstein database is the largest database in the field of organic chemistry. Since 2009, the content has been maintained and distributed by Elsevier Information Systems in Frankfurt under the name "Reaxys". The content covers more than 200 years of chemistry and has been abstracted from several thousands of journal titles, books and patents. Today the data is drawn from selected journals (400 titles) and chemistry patents, and the extraction process for each reaction or substance data included needs to meet three conditions:

1. It has a chemical structure
2. It is supported by an experimental fact (property, preparation, reaction)
3. It has a credible citation

Journals covered include *Advanced Synthesis and Catalysis, Angewandte Chemie , Journal of American Chemical Society, Journal of Organometallic Chemistry, Synlett* and *Tetrahedron.* [8][9]

## 6.2 Data Assessment

During my work with the Beilstain database have I found several issues with the data stored. First problem is that there are multiple instances of the same compound, but with different Reaxys IDs. As seen in fig.4 have I found four different IDs for the compound Dysidiolide. These were found when I tried to reproduce the different synthesis plans of Dysidiolide from [1] using the referenced articles where each synthesis plan origins. E. J. Corey's version of Dysidiolide have the ID 8171938, Boukouvalas have two different with ID 7601810 and 7910427 and Danishefsky have the ID 7910428.
Since we have four different IDs for Dysidiolide we can't state a single goal compound to our program that would result in giving us these three synthesis plans.



**Reaxys ID: 7601810**

$C_{25}H_{38}O_4$ 402.574 7601810

Identification | Bioactivity - 2 | Preparations - 16

Physical Data - 6 | Other Data - 1 | Reactions - 16

Spectra - 5 | Documents - 7

**Reaxys ID: 7910427**

$C_{25}H_{38}O_4$ 402.574 7910427

Identification | Preparations - 16

Physical Data - 1 | Reactions - 16

Documents - 1

**5-hydroxy-4-{1-hydroxy-2-[1,2,5-trimethyl-5-(4-methyl-pent-4-enyl)-1,2,3,5,6,7,8,8a-octahydro-naphthalen-1-yl]-ethyl}-5H-furan-2-one**

$C_{25}H_{38}O_4$ 402.574 7910428

Identification | Preparations - 52

Spectra - 3 | Reactions - 52

Documents - 3

**dysidiolide**

$C_{25}H_{38}O_4$ 402.574 8171938

Identification | Preparations - 24

Spectra - 4 | Reactions - 24

Bioactivity - 6 | Documents - 4

Fig. 4: Four different instances of Dysidiolide in Reaxys.

So what is causing this multiple ID issue? If we study fig. 4 can we see that the molecular weight and molecular formular are exactly the same, however if we look at the close up of the molecular structure in fig. 5 the compounds are not structured in the same way, even though the compound is the same. The main differences is:

1. Which way some of the substructures are facing. Example: The lower $C_4O_3H_4$ is rotated differently in each instance or that we write $H_2C$ instead of $CH_2$.

2. The bond between two chemical elements are notated.Example: The bond to the $OH$ in the bottom of the structure is either a "single" (d), "single down" (a) (c) or "single up or down" (b).

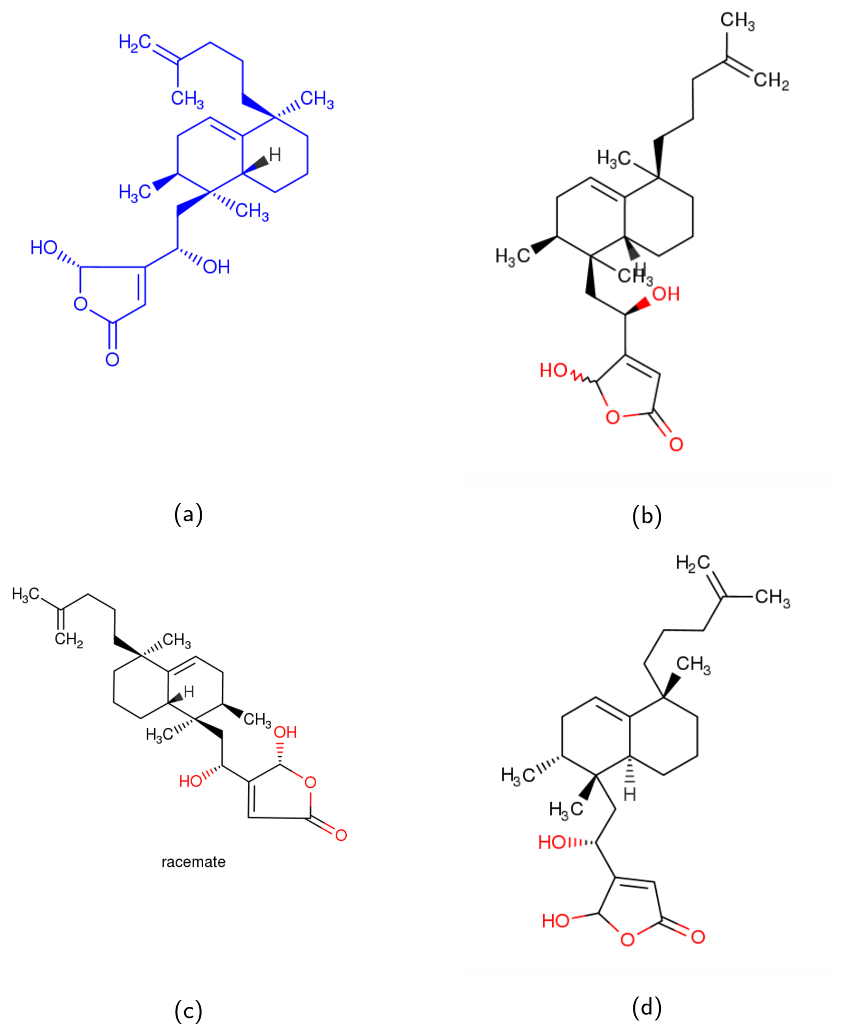
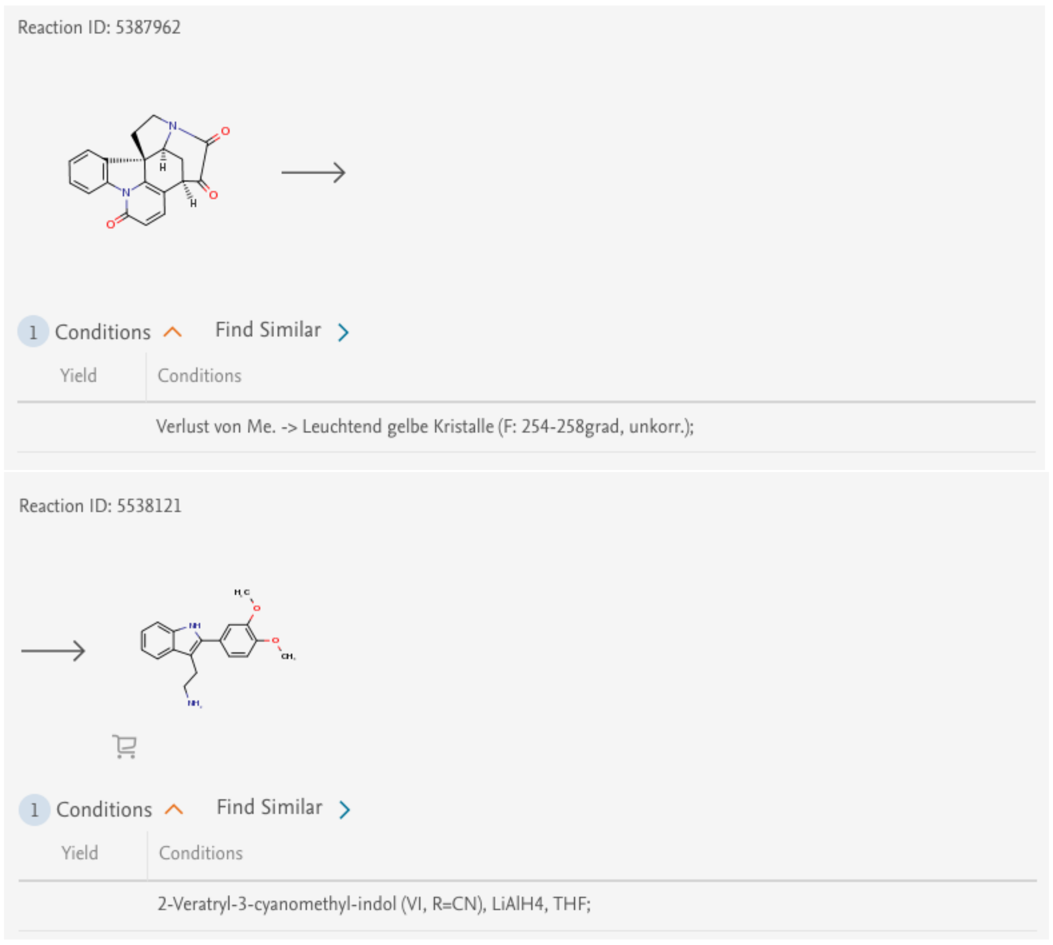


Fig. 5: Different versions of Dysidiolide: ID 7601810 (a), ID 7910427(b), ID 7910428(c), ID 8171938(d)

Benyt Dysidiolide til at vise at der er flere udgaver af samme compounds og at der er nogle som ikke hedder det som de er (Kun har REAXYSID eller et helt andet navn).
Find reaktioner som ikke har et educt eller et product



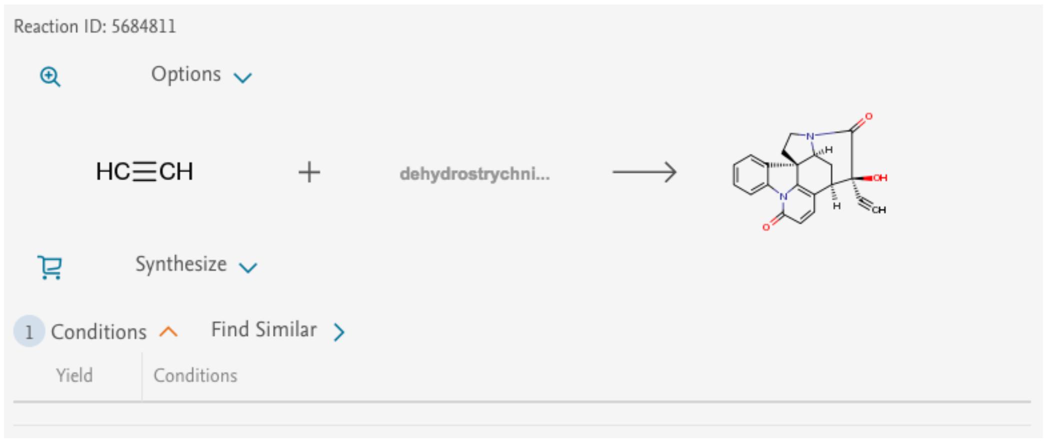
Reaction ID: 5387962

| Yield | Conditions |
|---|---|
| | Verlust von Me. -> Leuchtend gelbe Kristalle (F: 254-258grad, unkorr.); |

Reaction ID: 5538121

| Yield | Conditions |
|---|---|
| | 2-Veratryl-3-cyanomethyl-indol (VI, R=CN), LiAlH4, THF; |

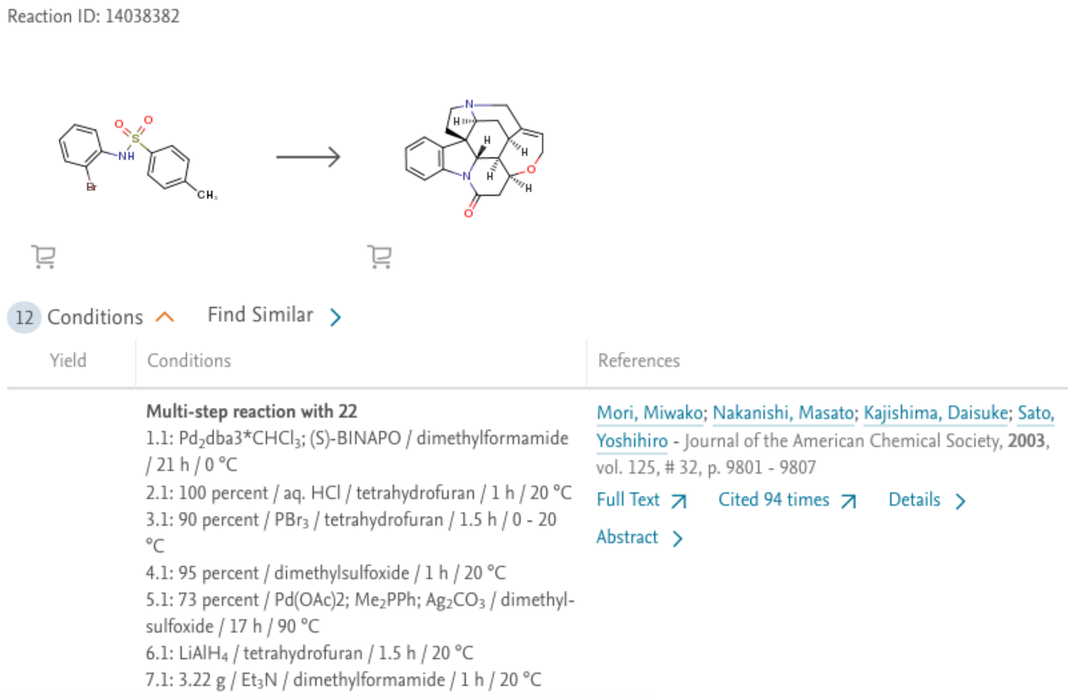
Fig. 6: Example of reactions that either is missing educts or products.



Reaction ID: 5684811

| Yield | Conditions |
|---|---|

Fig. 7: Example of an reaction with a missing educt.

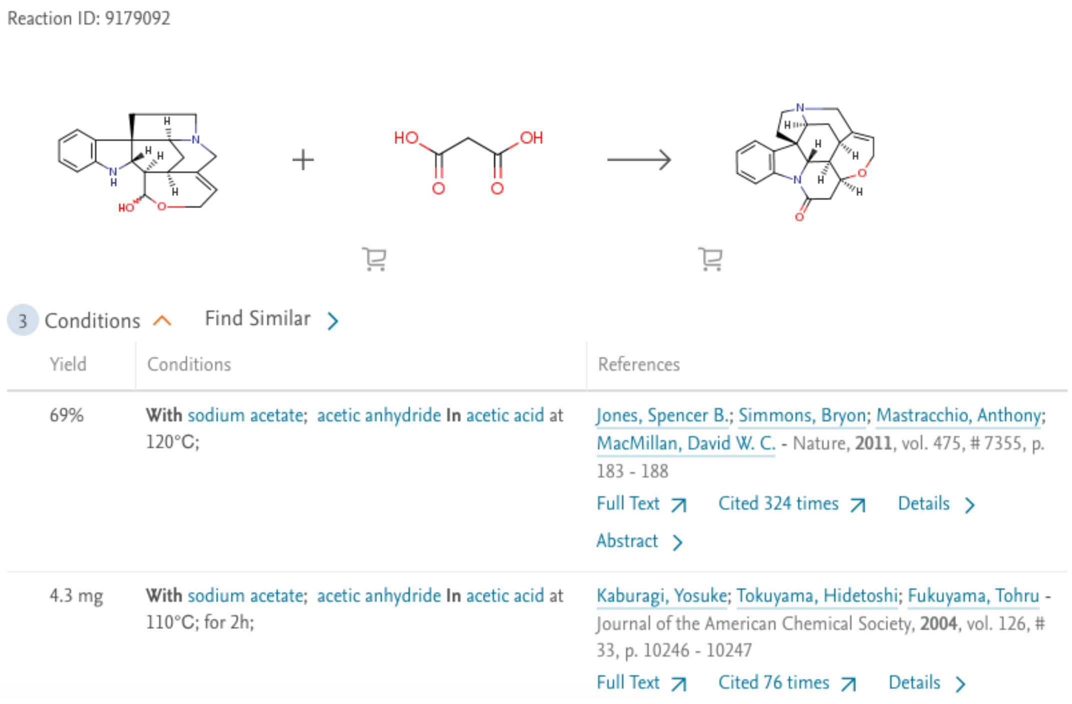


Reaction ID: 14038382

12 Conditions Find Similar

| Yield | Conditions | References |
|---|---|---|
| | **Multi-step reaction with 22**<br>1.1: $Pd_2dba3*CHCl_3$; (S)-BINAPO / dimethylformamide / 21 h / 0 °C<br>2.1: 100 percent / aq. HCl / tetrahydrofuran / 1 h / 20 °C<br>3.1: 90 percent / $PBr_3$ / tetrahydrofuran / 1.5 h / 0 - 20 °C<br>4.1: 95 percent / dimethylsulfoxide / 1 h / 20 °C<br>5.1: 73 percent / Pd(OAc)2; $Me_2PPh$; $Ag_2CO_3$ / dimethyl-sulfoxide / 17 h / 90 °C<br>6.1: $LiAlH_4$ / tetrahydrofuran / 1.5 h / 20 °C<br>7.1: 3.22 g / $Et_3N$ / dimethylformamide / 1 h / 20 °C | Mori, Miwako; Nakanishi, Masato; Kajishima, Daisuke; Sato, Yoshihiro - Journal of the American Chemical Society, **2003**, vol. 125, # 32, p. 9801 - 9807<br>Full Text Cited 94 times Details<br>Abstract |

Fig. 8: Example of a multistep reaction.



Reaction ID: 9179092

3 Conditions Find Similar

| Yield | Conditions | References |
|---|---|---|
| 69% | **With** sodium acetate; acetic anhydride **In** acetic acid at 120°C; | Jones, Spencer B.; Simmons, Bryon; Mastracchio, Anthony; MacMillan, David W. C. - Nature, **2011**, vol. 475, # 7355, p. 183 - 188<br>Full Text Cited 324 times Details<br>Abstract |
| 4.3 mg | **With** sodium acetate; acetic anhydride **In** acetic acid at 110°C; for 2h; | Kaburagi, Yosuke; Tokuyama, Hidetoshi; Fukuyama, Tohru - Journal of the American Chemical Society, **2004**, vol. 126, # 33, p. 10246 - 10247<br>Full Text Cited 76 times Details |

Fig. 9: Example of inconsistent yield notation.

## 6.3 Making A B-Hypergraph

## 6.4 Testing

### 6.4.1 Strychnine

### 6.4.2 Compound 2

### 6.4.3 Compound 3

kkkk [1]

## 7 Konklusion

## Books

[1] R. W. Hoffmann, *Elements of Synthesis Planning.* Springer Berlin Heidelberg, 2009.

## Articles

[2] S. Szymkuc, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk, and B. A. Grzybowski, “Computer-assisted synthetic planning: The end of the beginning”, *Angewandte Chemie International Edition*, no. 55, pp. 5904–5937, 2016.

[3] R. Fagerberg, C. Flamm, R. Kianian, D. Merkle, and P. F. Stadler, “Finding the K best synthesis plans”, 2017, Unpublished Article.

[4] J. Y. Yen, “Finding the K shortest loopless paths in a network”, *Management Science*, vol. 17, no. 11, pp. 712–716, Jul. 1971.

[5] L. R. Nielsen, K. A. Andersen, and D. Pretolani, “Finding the K shortest hyperpaths: Algorithms and applications”, 2002.

## Other

[6] C. G. Lützen and D. F. Johansen, “A computational and mathematical approach to synthesis planning”, Master’s thesis, University of Southern Denmark, 2015.

[7] Sep. 2017. [Online]. Available: `http://en.cppreference.com/w/cpp/container/vector_bool`.

[8] Oct. 2017. [Online]. Available: `https://en.wikipedia.org/wiki/Reaxys`.

[9] Oct. 2017. [Online]. Available: `https://en.wikipedia.org/wiki/Beilstein_database`.

## 8 Appendiks