RETROSYNTHETIC ANALYSIS & AI

(https://doi.org/10.6084/m9.figshare.22730582.v2)

Introduction

Recently much has been heard about artificial intelligence and the risks of using this technology. Therefore, we decided to use ChatGPT Mar23Version to write part of this introduction.

Retrosynthetic analysis as defined by Elias Corey (ChatGPT answer)

Retrosynthetic analysis is a powerful problem-solving approach in organic chemistry that was pioneered by Elias Corey in the 1960s. The idea behind retrosynthetic analysis is to break down complex molecules into simpler, readily available starting materials through a series of retrosynthetic steps. This approach helps chemists plan a synthesis route by working backward from the target molecule to identify the necessary starting materials and reactions needed to make it.

Corey's approach to retrosynthetic analysis involves identifying key functional groups in the target molecule and proposing ways to disconnect them to form simpler, more easily accessible building blocks. The goal is to break down the target molecule into a series of intermediates that can be synthesized from simpler starting materials using known chemical reactions.

To illustrate this approach, consider the synthesis of ibuprofen, a common pain reliever. The target molecule contains an isobutyl group attached to a phenyl ring, which can be disconnected at the carbon-carbon bond to form two simpler building blocks: isobutyl bromide and benzene. These building blocks can then be further broken down into even simpler starting materials, such as isobutylene and hydrobromic acid for the isobutyl bromide, and benzene and ethylene for the phenyl ring.

Overall, Corey's retrosynthetic analysis provides a powerful tool for planning organic synthesis routes by breaking down complex molecules into simpler, more readily available building blocks.

in & out science

AI-based retrosynthetic analysis of organic molecules (ChatGPT answer)

In recent years, artificial intelligence (AI) has been increasingly applied to retrosynthetic analysis to help chemists predict the most efficient routes for synthesizing organic molecules. AI-based retrosynthetic analysis uses machine learning algorithms trained on large datasets of known chemical reactions to identify the most likely reaction pathways for synthesizing a target molecule.

One popular approach is to use a neural network trained on reaction databases to predict the retrosynthetic steps for a target molecule. The neural network is fed information about the target molecule's structure, functional groups, and reaction conditions, and it uses this information to predict the most likely retrosynthetic steps leading to the target molecule. The predicted steps are then ranked based on their likelihood, and the chemist can choose the most promising route for synthesis.

Another approach is to use a knowledge-based system that relies on rules and heuristics to guide the retrosynthetic analysis process. These systems typically use a database of known reactions and synthetic strategies to propose possible routes for synthesizing a target molecule. The chemist can then use this information to select the most promising route for synthesis.

Overall, AI-based retrosynthetic analysis has the potential to greatly accelerate the drug discovery process and improve the efficiency of organic synthesis. By providing chemists with more accurate and efficient tools for planning synthetic routes, AI-based retrosynthetic analysis can help to reduce the time and resources required to develop new drugs and other important organic compounds.

The texts originated were not too bad. However, analyzing the answers one finds few things worth to mention. The first “odd” thing is the ibuprofen example in the Corey’s definition of retrosynthetic analysis. Although a sort of valid example, the molecule of ibuprofen contains more than an isobutyl group and a benzene ring (Figure 1).

in & out science

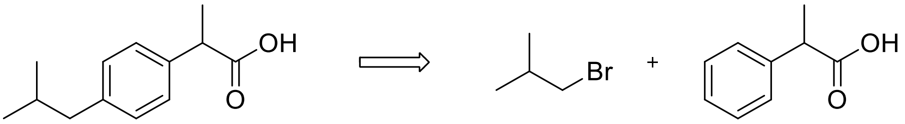


Figure 1. Ibuprofen example corrected.

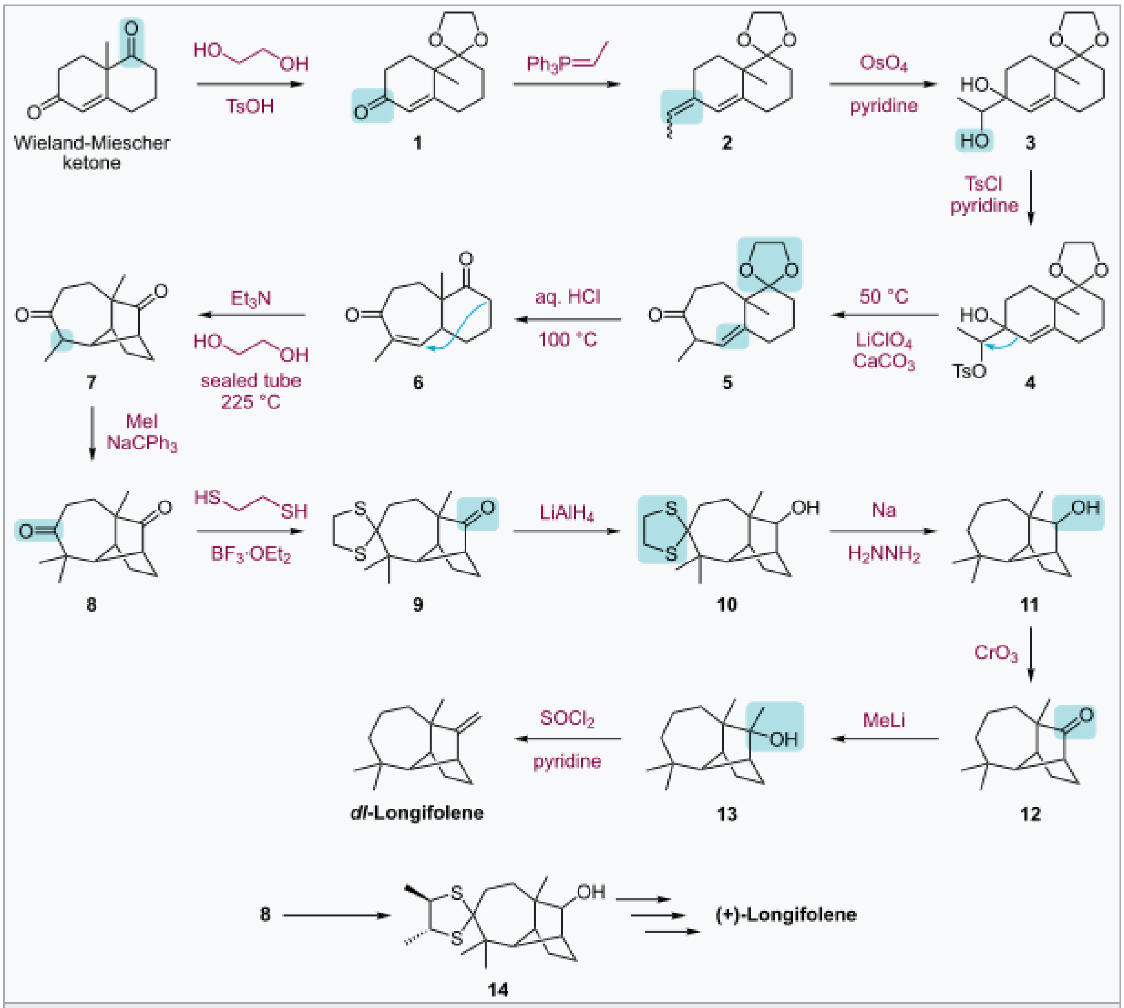
The second interesting statement from ChatGPT is that retrosynthetic analysis using artificial intelligence bases its knowledge on *"…large datasets of known chemical reactions…"*, but what about the unknown. This argument could go on and on, but the fact is that AI does not seem "to invent" new synthetic routes, instead, AI can be seen as an innovative approach to the retrosynthetic analysis of organic molecules using "available knowledge".

Results and Discussion

One way to put in perspective the argument in the previous paragraph, was to run a case example with an old molecule to see if the new AI-based retrosynthetic tools can generate unknown synthetic pathways. The molecule of choice was Longifolene, first synthesized in 1961 by Elias J. Corey (Figure 2).

Since Corey's ingenious way to synthesize Longifolene, there has been a number of new approaches to the synthesis of this molecule (https://chem.libretexts.org/Bookshelves/Organic_Chemistry/Logic_of_Organic_Synthesis_(Rao)/07%3A_Strategies_in_Longfolene_Synthesis).

in & out science



Longifolene total synthesis by Corey.svg

Figure 2. Longifolene structure (racemic) and Corey's synthetic approach.

The different approaches so far reported, have something in common, all of them use a common retrosynthesis strategy, the introduction of functional groups that enable different C-C bond forming reactions.

This strategy (functional group introduction) has been implemented in some of the systems tested for this article. However, none of them came close to give a first (or second) disconnection that could be considered as new approach to the synthesis of Longifolene.

Without getting into details about the different answers obtained, the systems used for this study were: ASKCOS, AiZynthFinder, ChemPlanner, ICSynth and IBM RXN for Chemistry. The newer systems (ASKCOS, AiZynthFinder and IBM RXN for Chemistry) have, besides the commercial version, an open access version that probably is not as capable.

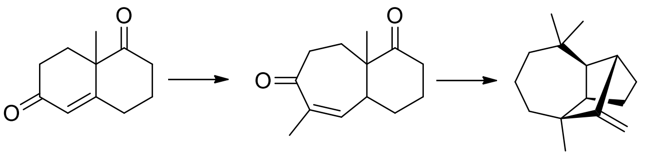
It's worth to remark at this point that the task at hand (find new synthesis for Longifolene) was not an easy task due to the number of different approaches and synthesis of this molecule over the years. Lei, B. et al reported in 1993 a comprehensive article summarizing the different retrosynthetic analysis done by academic groups (https://doi.org/10.1021/JO00060A039) (Figure 3).

Longifolene was then used as target molecule in each of the systems mentioned above (using mostly non-commercial versions when available). Regardless that all systems gave, more or less valid answers, none of them gave new ideas for the synthesis of Longifolene.

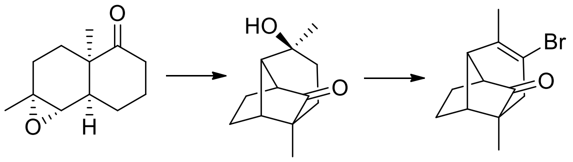
Typical retrosynthetic strategies such as functional group introduction and complexity reduction (mainly by reducing the number of rings) were observed in all systems used. However, some of the systems, focused too much in finding similar reactions in the literature to back up a given suggestion and missed the opportunity to introduce new synthetic approaches. Others kept modifying the functional groups introduced in the previous step without simplifying the synthetic target, something that represents the essence of the retrosynthetic analysis (see Corey's definition in the intro).

Ideally, the perfect retrosynthetic analysis system would it be a combination of all of them, a utopia in today's world where everyone knows better, and economical departments runs science.

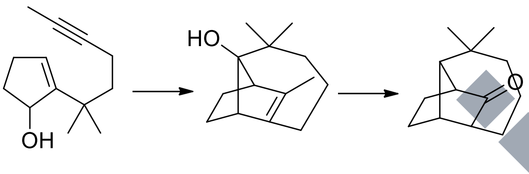
in & out science



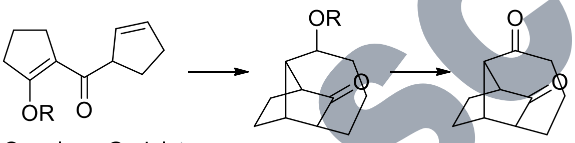
Corey, Ohno, Mitra, Vatakencherry



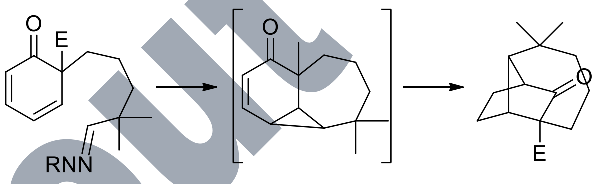
McMurry, Isser



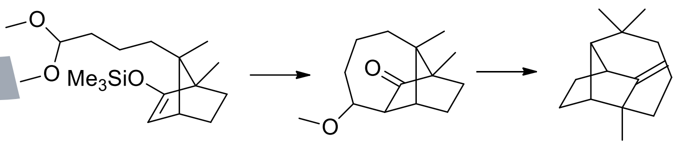
Johnson, Volkmann, Andrewys



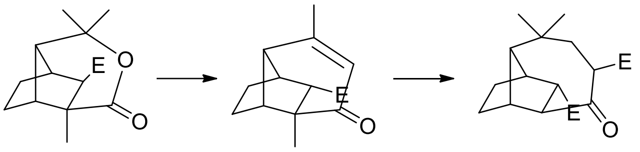
Oppolzer, Godel



Shultz, Puig



Money, Kuo



Liu, Ho, Yeh, Yule

Figure 3. (1) Corey, E. J. Ohno, M.; Mitra, R. B.; Vatakencherry, P. A. J. Am. Chem. Soc. 1964,86,478. (2) McMurry, J. E.; Isser, S. J. J. Am. Chem. SOC. 1972, 94, 7132. (3) Volkmann, R. A.; Andrews, G. C.; Johnson, W. S. J. Am. Chem. Soc. 1975, 97,4777. (4) Oppolzer, W.; Godel, T. Helo. Chim. Acta 1984, 67, 1154. (5) Schultz, A. G.; Puig, 5. J. Org. Chem. 1985, 50, 915. (6) Kuo, D. L.; Money, T. Can. J. Chem. 1988,66,1794. (7) Ho, T. L.; Yeh, W. L.; Yule, J.; Liu, H.-J. Can. J. Chem. 1992,70, 1375.

in & out science

Having worked with Computer Aided Synthesis Design (CASD) systems since 2009 made for high expectations when testing CASD systems based on artificial intelligence. However, besides the argument that the new systems give more reliable suggestions, the fact was that the retrosynthetic approach suggested by any of the CASD systems used was not enough to come up with a brand-new synthetic route.

The goal with this study was never to compare commercially available systems and rank them since all of them have interesting features and at the end, the idea generation process depends mainly on the eyes of the chemist looking at the suggestions.

As scientific hint of all it has been discussed in this article, it is the fact that none of the CASD systems suggested a ring-closing metathesis transformation to generate the seven-membered ring in the Longifolene.[1]

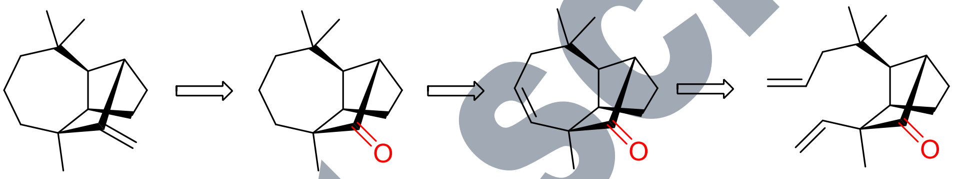


Figure 4. Suggested new synthesis of Longifolene using ring-closing metathesis as key step. NOTE that the ketone might need functional group protection for this suggestion to work.

The synthetic suggestion depicted in Figure 4 might or might not be synthetically feasible, but in any case, the introduction of functional groups such the ketone and the double bond were suggested by some of the CASD systems tested. However, none of them suggested what in the head of many synthetic chemists would be a natural synthetic step.

Without getting too much into chemistry and synthetic details, the example presented here showed that in route design, AI is not (yet) at a level where we can blindly follow its results. In & Out Science uses a combination of AI based retrosynthetic analysis systems plus data mining techniques to find new synthesis of complex organic molecules.[2]

Longifolene New Route Design (Hypothetically)

in & out science

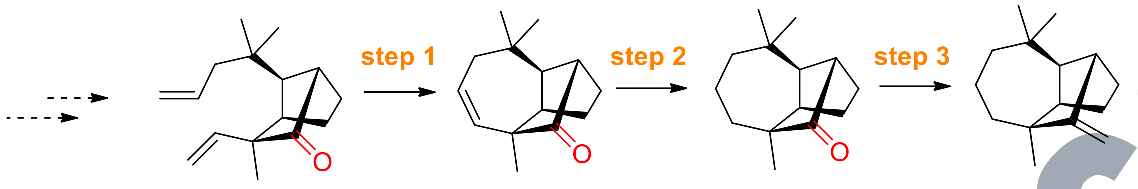
---

[1] To the best of our knowledge, as today, no total synthesis of Longifolene has been published using ring closing metathesis in any of the synthetic steps.

[2] **The power of data mining in route design: Application to the first synthesis of a biindole alkaloid from the fruiting bodies of Tricholoma sciodes**, *Manuscript in preparation.*

The next important question to be answered would be: Would, a new synthesis of Longifolene based on the disconnections made in Figure 4, work?

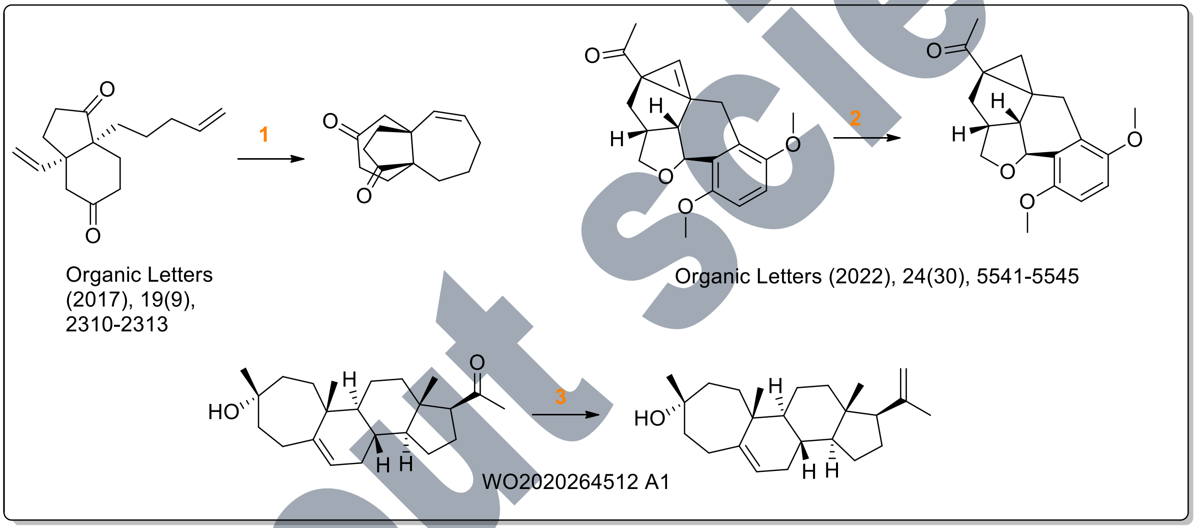
In & Out Science looked for precedents in the literature (using data mining techniques) for the synthetic steps in Figure 5.



4 **3** **2** **1**

*Figure 5. Considering that compound **4** is synthetically feasible, steps 1, 2 and 3 would represent the last steps in the new route design for Longifolene.*

Some of the precedents found:



Organic Letters (2017), 19(9), 2310-2313

Organic Letters (2022), 24(30), 5541-5545

WO2020264512 A1

At this point it would be impossible to answer the question, however, looking at the precedents it could.

Disclaimer: The CASD systems here tested were used following the default settings from the providers in their open-source version. Some of them did not run as far as three steps, but none of the one-step suggestions were compatible with the retrosynthetic path depicted in Figure 4.

no preaching, no teaching, just a

perspective and an opinion

in & out science