COMPUTER AIDED SYNTHESIS DESIGN AND TOOLS FOR SYNTHESIS PLANNING

CONTENT

FOREWORD

This course is planned for students that want an introduction on cheminformatics and in particular on computer aided synthesis design (CASD). During the duration of the course, several new topics will be introduced to the students, most of these new topics will lie in the category “good to know information”, and a deeper knowledge will not be required for the porpoise of the course.

Nevertheless, the student will be provided with sufficient references to dig deeper into this knowledge area, if there so desired.

1. INTRODUCTION

The first question when presenting software for retrosynthetic analysis is, do we really need it?

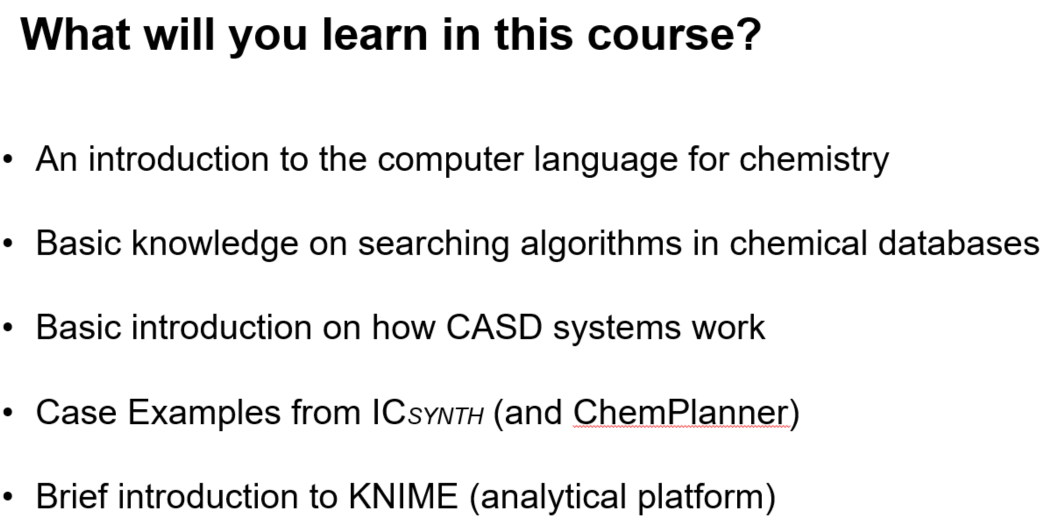
All chemists think that the process of generating new synthetic ideas for the synthesis of complex molecules is one of the most rewarding and fun tasks to do in organic synthetic chemistry, so, why should a computer get all the fun?

Thankfully, Computer Assisted Synthesis Design (CASD) systems help us to increase our knowledge-base and to open up to new ways of retrosynthetic analysis. We, the chemists, will still have all the fun solving complex synthetic problems and the good news is that now we have even more sophisticated “toys” to succeed with our synthetic problems.

The second debate is about the name CASD. There are many other ways to name systems that help chemists to solve synthetic problems, *computer aided synthesis design* (also CASD), *computer assisted organic synthesis* (CAOS), *organic chemical simulation of synthesis*

(OCSS), *computer assisted/aided synthesis planning* (CASP), etc. Aside from the name that one choses to use, the bottom line is that all of them do exactly the same, that is, the use of computers to expand our synthetic knowledge within a specific synthetic problem.

During the 2 hours of this course the student will get an introduction to the computer language for chemistry, although in very basic bases. The most important part of the course is to make the student understand how computers “think” about chemistry, and therefore to understand how computers can help chemists with their synthetic problems.



What will you learn in this course?

- An introduction to the computer language for chemistry
- Basic knowledge on searching algorithms in chemical databases
- Basic introduction on how CASD systems work
- Case Examples from IC*SYNTH* (and ChemPlanner)
- Brief introduction to KNIME (analytical platform)

1. INTRODUCTION

2. THE LOGIC OF CHEMICAL SYNTHESIS (THE ORIGIN OF COMPUTER AIDED SYNTHESIS PLANNING)[1]

Corey was the first showing us a synthetic tree (Figure 1).

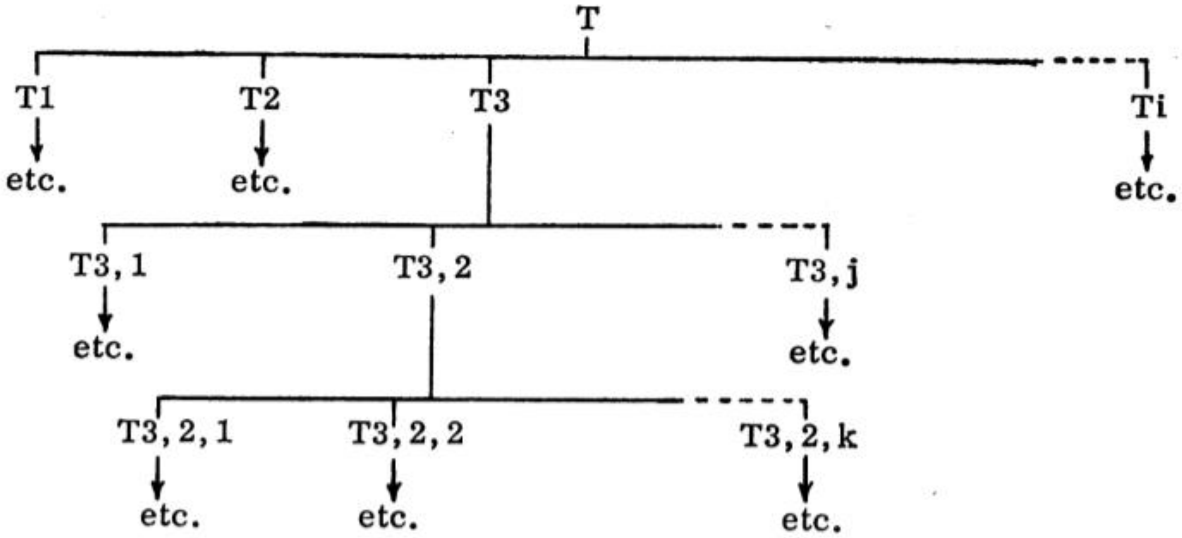


Fig. 1. Synthetic analysis of target *T* generates a "tree" of intermediate precursor structures.

Figure 1. A synthetic tree first suggested by Corey and Wipke.[1]

He was also the first chemist who believed in the power of computer science. Back in 1969 he already suggested the processes (workflow) involved in a computer assisted synthesis planning named OCSS (Organic Chemical Synthesis Simulation, Figure 2).

OCSS was also the first system ever using interactive graphics for the input of target molecules.

[1] Corey, E. J. and Wipke, W. T. Science 1969, 166, 178-192.

For a recommended review about CASD[2] see, Ravitz, Orr et al. *WIREs Comput Mol Sci* **2012,** *2*: 79–107. doi: 10.1002/wcms.61.

[2] For more reviews see; **a)** Feng F, Lai L and Pei J, *Front. Chem.* (2018), 6:199. doi: 10.3389/fchem.2018.00199. **b)** Ugi, I., Bauer, J., Bley, K., Dengler, A., Dietz, A., Fontain, E., … Stein, N. *Angewandte Chemie International Edition in English*, (1993), *32*(2), 201–227. doi.org/10.1002/anie.199302011. **c)** de Almeida, A. F., Moreira, R., & Rodrigues, T. *Nature Reviews Chemistry*, (2019), *3*(10), 589–604. doi.org/10.1038/s41570-019-0124-0. d) Feng, F., Lai, L., & Pei, J. *Frontiers in Chemistry*, (2018), *6*(JUN). https://doi.org/10.3389/fchem.2018.00199. **d)** Warr, W. A. *Molecular Informatics*, (2014), *33*(6–7), 469–476. https://doi.org/10.1002/minf.201400052

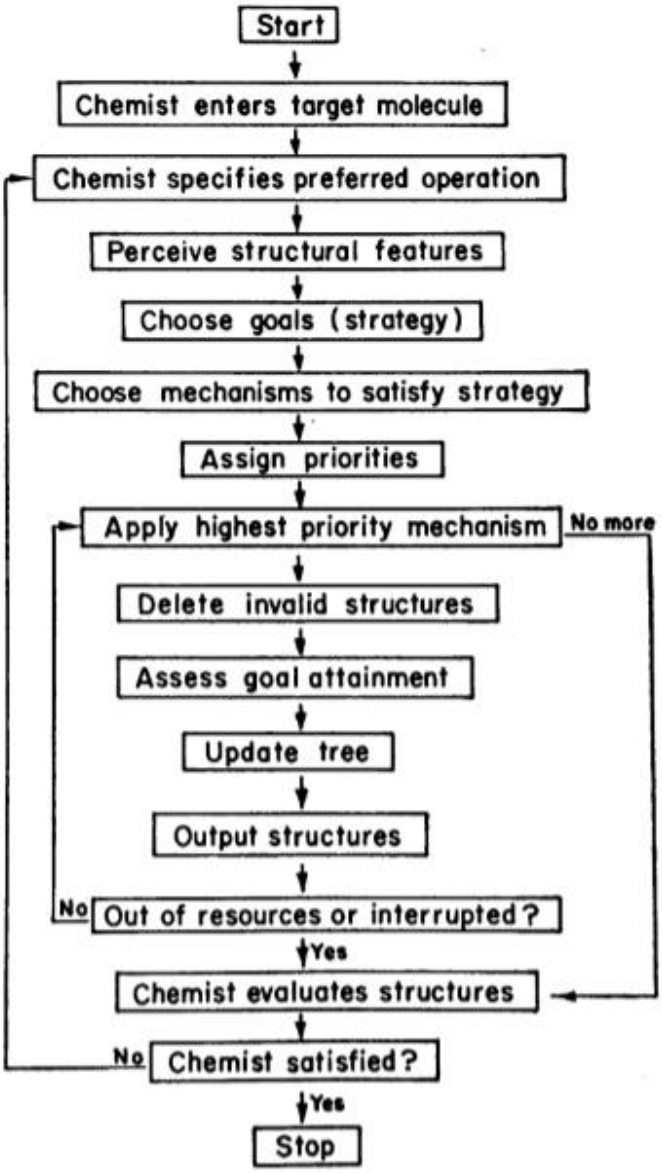

Start
Chemist enters target molecule
Chemist specifies preferred operation
Perceive structural features
Choose goals (strategy)
Choose mechanisms to satisfy strategy
Assign priorities
Apply highest priority mechanism
No more
Delete invalid structures
Assess goal attainment
Update tree
Output structures
No
Out of resources or interrupted?
Yes
Chemist evaluates structures
No
Chemist satisfied?
Yes
Stop

Fig. 2. Processing scheme for machine-assisted logic-centered synthetic analysis.

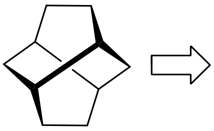
Figure 2. Corey first suggestion for a computer assisted organic synthesis process.

One key aspect of Corey's system was the introduction of heuristics to help with synthetic problems.

Why heuristics?

This question is best answered with an example.

What bond would you disconnect first?



Heuristic Rule #. Disconnect bonds that simplified the symmetry of the molecule.

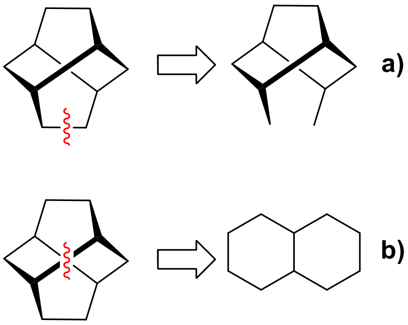


Figure 3. Disconnection b) will be the most appropriate.

Heuristics help us to simplify a synthetic problem.

Computers need a set of rules to analyze complex problems. These rules are the “thinking” part of the computer.

Heuristics Definition (Wikipedia)

https://en.wikipedia.org/wiki/Heuristic

A heuristic technique (/hjʊəˈrɪstɪk/; Ancient Greek: εὑρίσκω, "find" or "discover"), or a heuristic for short, is any approach to problem solving or self-discovery that employs a practical method that is not guaranteed to be optimal, perfect or rational, but which is nevertheless sufficient for reaching an immediate, short-term goal.

3. CHEMICAL INFORMATION SYSTEMS (CIS)

A CIS is a collection of computerized data storage, and retrieval components for chemical information, covering such aspects as structure, names, spectra, toxicology, hazards, literature references and molecular modeling. Each component is essentially a 'standalone' system, but they share utility software and therefore composite searches within different databases are easily performed.[3]

Edward A. Feigenbaum is consider the father of expert systems for his work back in the 60's. It was about that time when the term cheminformatics was first used.[4]

Cheminformatics is usually defined in terms of the application of computer science and information technology to problems in the chemical sciences. Brown[5] introduced the term chemoinformatics in 1998, in the context of drug discovery, although informatics techniques have been applied in chemistry since 1950s, and cheminformatics now relates to a broader set of contexts.

The presentation in the link below shows some aspects of the cheminformatics, such as:

- name to structure (or vice versa).
- structure to smiles
- structure to molfile
- etc…

It overlaps with some content in this course but not much.

https://www.slideshare.net/baoilleach/cheminformatics-13581857

[3] Boyle, L. (1986). *The chemical information system. TrAC Trends in Analytical Chemistry, 5(10), VI–VII.* doi:10.1016/0165-9936(86)85069-5

[4] Alan M. DuffieldAlexander V. RobertsonCarl DjerassiBruce G. BuchananGeorgia L. SutherlandEdward A. FeigenbaumJoshua Lederberg, *J. Am. Chem. Soc.* 1969, 91, 11, 2977-2981. https://doi.org/10.1021/ja01039a026.

[5] Brown FK. Chemoinformatics: what is it and how does it impact drug discovery. *Annu Rep Med Chem*. 1998;33:375–384.

a. Reaction representation (smiles, inchi, smarts, smirks, molfiles, etc…)

One important aspect in cheminformatics is how to store structural information in a computer. The main problem is that although computer now can recognize 3D images, this process is slow and it will take forever to revise millions of structures in a database, so, the solution is to find another way to keep the structural information and be able to retrieve it as fast as possible.

Some of the most common ways to store structural information in databases are described below, together with some links for more information.

Good general overview @ https://chem.libretexts.org/Courses/University_of_Arkansas_Little_Rock/ChemInformatics_(2017)%3A_Chem_4399%2F%2F5399/2.3%3A_Chemical_Representations_on_Computer%3A_Part_III

Smiles

SMILES (Simplified Molecular Input Line Entry System) is a chemical notation that allows a user to represent a chemical structure in a way that can be used by the computer. ...

- https://archive.epa.gov/med/med_archive_03/web/html/smiles.html
- https://www.daylight.com/dayhtml_tutorials/languages/smiles/index.html

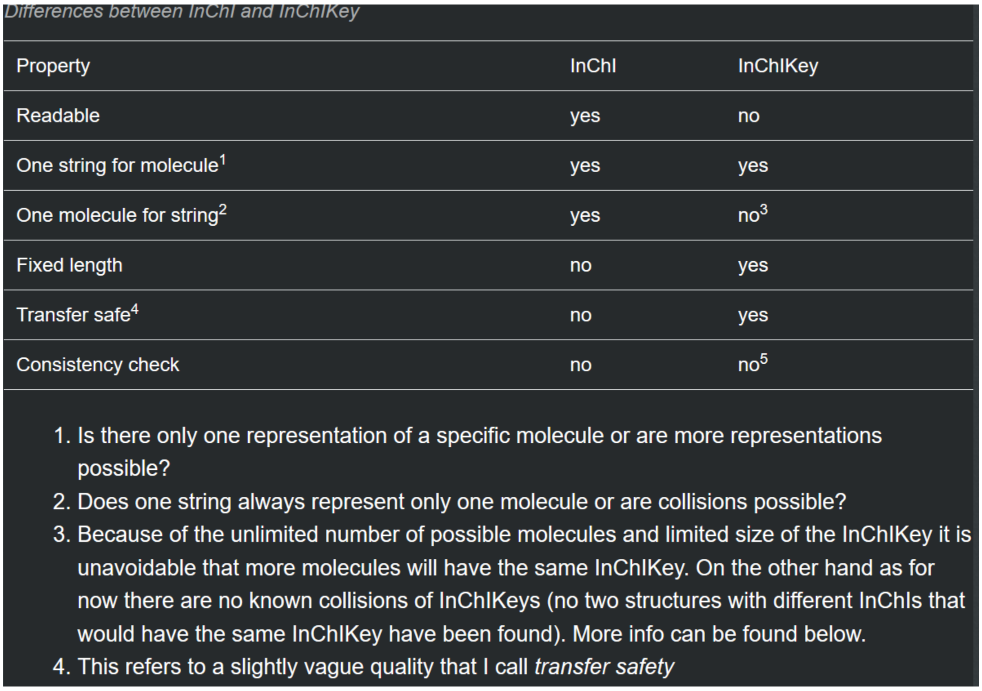
InChi

The IUPAC International Chemical Identifier (InChI^{TM}) is a non-proprietary identifier for chemical substances that can be used in printed and electronic data sources thus enabling easier linking of diverse data compilations.

https://jcheminf.biomedcentral.com/track/pdf/10.1186/s13321-015-0068-4

InChiKey

https://inchi.info/inchikey_overview_en.html



Differences between InChI and InChIKey

Property	InChI	InChIKey
Readable	yes	no
One string for molecule[1]	yes	yes
One molecule for string[2]	yes	no[3]
Fixed length	no	yes
Transfer safe[4]	no	yes
Consistency check	no	no[5]

1. Is there only one representation of a specific molecule or are more representations possible?
2. Does one string always represent only one molecule or are collisions possible?
3. Because of the unlimited number of possible molecules and limited size of the InChIKey it is unavoidable that more molecules will have the same InChIKey. On the other hand as for now there are no known collisions of InChIKeys (no two structures with different InChIs that would have the same InChIKey have been found). More info can be found below.
4. This refers to a slightly vague quality that I call *transfer safety*

Smarts:

SMILES arbitrary target specification (SMARTS) is a language for specifying substructural patterns in molecules. The SMARTS line notation is expressive and allows extremely precise and transparent substructural specification and atom typing.

https://docs.eyesopen.com/toolkits/python/oechemtk/SMARTS.html

https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html

http://organica1.org/seminario/daylight.pdf

(FF) => Best option to define substructure queries!!

Smirks:

Representation of generic transformations. Very related to Smiles and Smarts.

https://www.daylight.com/dayhtml/doc/theory/theory.smirks.html

Molfile:

http://c4.cabrillo.edu/404/ctfile.pdf => over 100 pages of explanation about connection tables, molfiles and sdfiles.

https://chem.libretexts.org/Courses/University_of_Arkansas_Little_Rock/ChemInformatics_(2017)%3A_Chem_4399%2F%2F5399/2.2%3A_Chemical_Representations_on_Computer%3A_Part_II/2.2.2%3A_Anatomy_of_a_MOL_file => More condense explanation about molfiles.

SDfiles

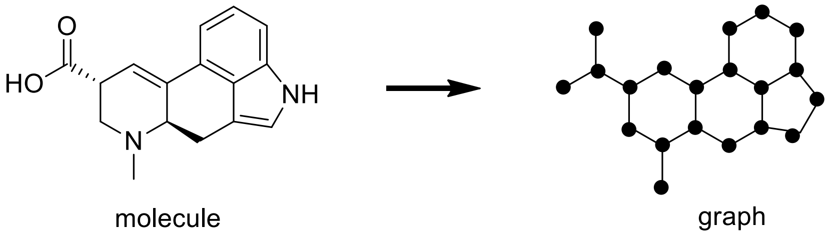
The main difference with the Molfile is that in a SDfile there is also molecular data associated. The structure of a SDfile can be seen in http://c4.cabrillo.edu/404/ctfile.pdf

b. Reaction database structure and searching algorithms (subgraph isomorphism, reaction mapping, reaction classification, fingerprints, etc…)

(PDF)\Thesis_SearchEngine4ChemicalDB_WangHao.pdf"
https://pdfs.semanticscholar.org/b728/294c24bc869d8fe4d4174d6fa2e39078b597.pdf

The usual way of storing unique structures in a computer is a canonicalized connection table: a listing of atoms and bonds, and other data, in tabular form. The structures can then be searched by substructure, that is, all the molecules in the database that contain a specified substructure can be identified.[6]

Structure searches are carried out by treating the structure as a graph,[7] with the atoms as nodes and the bonds as edges joining the nodes, and then by applying graph theory algorithms to carry out the match.



Subgraph Isomorphism

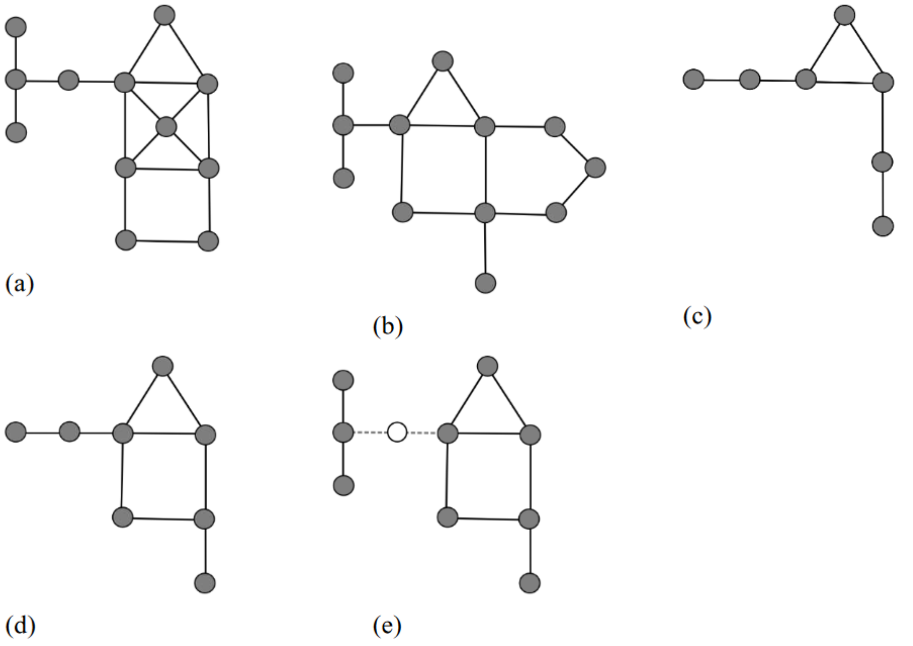


Figure 1: (a) and (b) represent the graphs G_1 and G_2. (c), (d) and (e) are respectively the MCIS, the cMCES and the dMCES for G_1 and G_2 (the white node in (e) is a feature from G_1, and has been included for ease of understanding but is not part of the dMCES).

[6] Barnard, J. M. (1993). Substructure Searching Methods: Old and New. Journal of Chemical Information and Computer Sciences, 33(4), 532–538. https://doi.org/10.1021/ci00014a001.

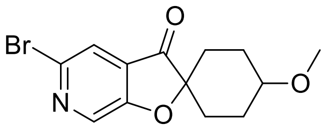
[7] For information about the origin of graph theory see: https://www.encyclopedia.com/science/encyclopedias-almanacs-transcripts-and-maps/birth-graph-theory-leonhard-euler-and-konigsberg-bridge-problem.

http://eprints.whiterose.ac.uk/102232/3/MCS_review_final.pdf

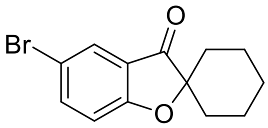
Originally, graph theory applied to chemical structure search was most commonly used in Medicinal Chemistry where ”the shape” of the molecule is an important factor for its biological activity.

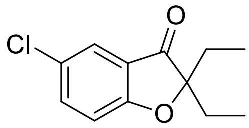
In CASD systems it has been used to find new ”ideas” for the synthesis of target molecules.

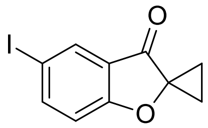
Example: Target molecule *(very simplified explanation)*

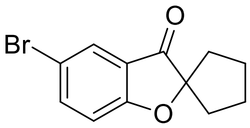


In order to find suitable disconnections and suggest reactions to make the disconnected bonds, a CASD system will look into all possible “graph” variations (see below) for suitable reactions.

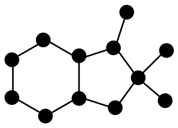








The common graph from all these structures is:



This kind of subgraph search corresponds with just one of the multiple processes that operate simultaneously at the backend of a CASD system.

Fingerprints

Molecular fingerprints are a way of encoding the structure of a molecule. The most common type of fingerprint is a series of binary digits (bits) that represent the presence or absence of particular substructures in the molecule. Comparing fingerprints allows you to determine the similarity between two molecules, to find matches to a query substructure, etc.[8]

https://towardsdatascience.com/a-practical-introduction-to-the-use-of-molecular-fingerprints-in-drug-discovery-7f15021be2b1

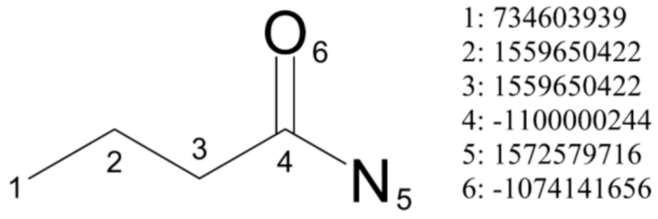
PROCESS TO GENERATE FP's

1. Assign each atom with an identifier

We choose an atom in the molecule and take note of:

- number of nearest-neighbor non-hydrogen atoms
- number of bonds attached to the atom (not including bonds to hydrogens)
- atomic number
- atomic mass
- number of hydrogens connected to the atom
- is the atom in a ring (1) or not (0)?

These values are hashed into an integer number. This process is repeated for each atom until all atoms have been assigned a hashed integer value. As an example, this is the result of this process on this arbitrary molecule:

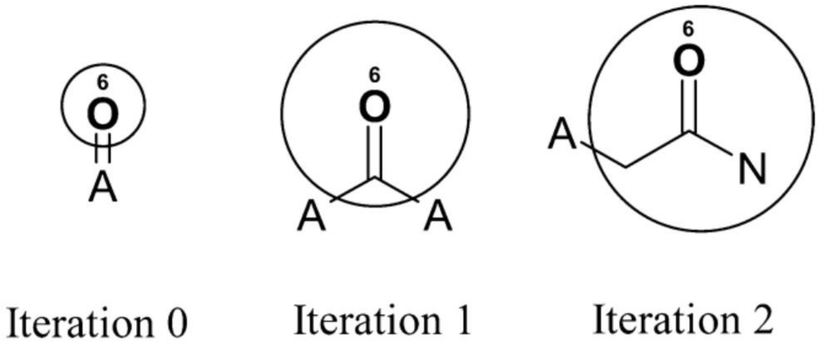


1: 734603939
2: 1559650422
3: 1559650422
4: -1100000244
5: 1572579716
6: -1074141656

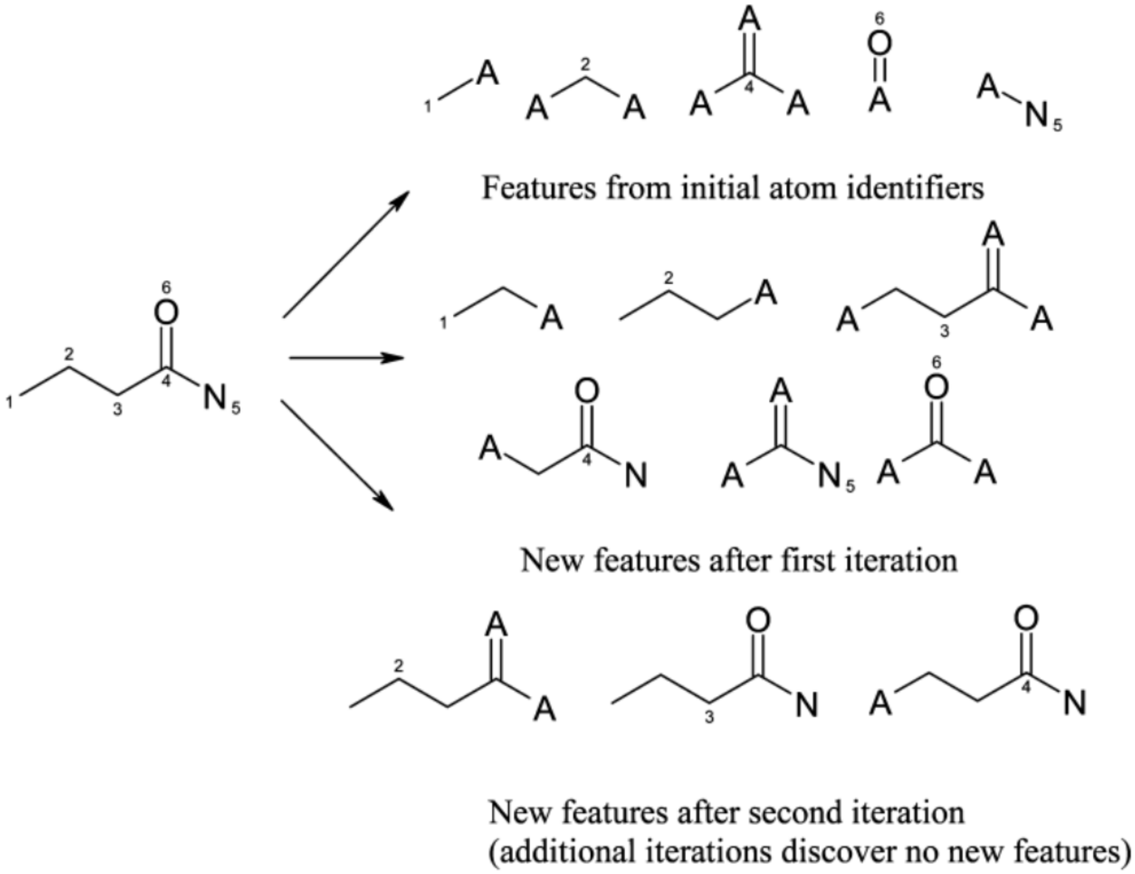
The first iteration

2. Update the identifiers of each atom, iteratively

[8] https://openbabel.org/docs/dev/Fingerprints/intro.html



3. Duplicate substructure removal



Features from initial atom identifiers

New features after first iteration

New features after second iteration
(additional iterations discover no new features)

4. Wrapping up

Finally, after the desired number of iterations are performed (2–4 for most purposes), we create an array of each atom identifier from each iteration level, having removed duplicates, and fold it into a length 2048 bit vector[9] using a hashing algorithm.

[9] Commonly used bit vector length, but can be as large as 16k or more.

For info on how it works a binary number system see:

https://www.mathsisfun.com/binary-number-system.html

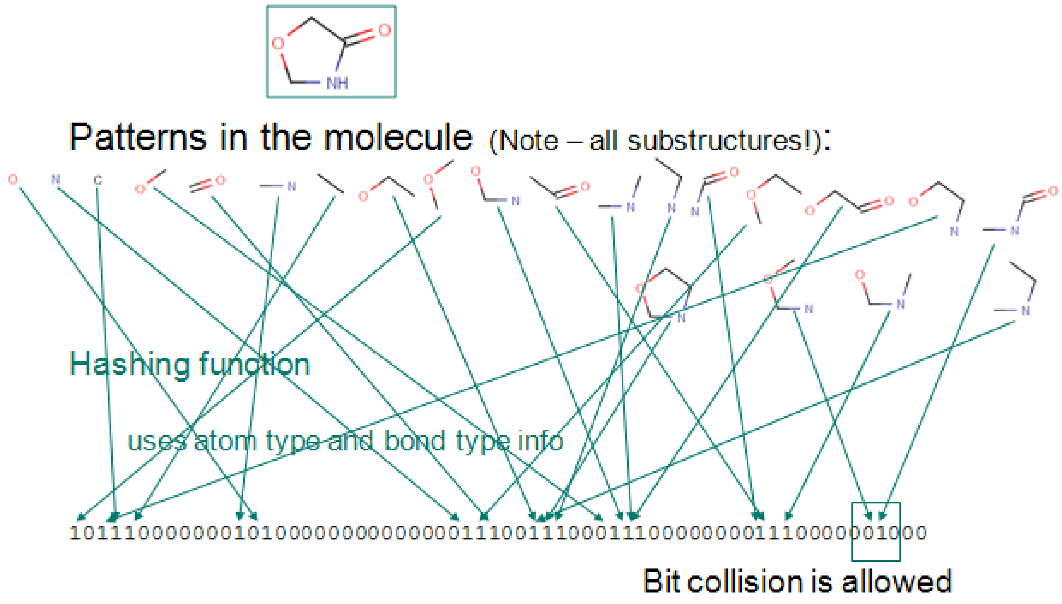


Figure 4. At the end of the day a fingerprint is not more than one huge binary vector that represents fragments of a molecule.

Reaction Classification, Reaction Center and Atom Mapping

For an automated system, before starting any retrosynthetic analysis, they must learn ”reactions”.

Computers uses different approach to classify reactions and to detect reaction centers compare with ”traditional chemists”.

Before a reaction can be classified, a computer must identify the reaction center, therefore, an atom mapping following the fate of the different atoms from starting materials to the final product must be in place. Different approaches are described in the literature but besides their differences, all of them do the same job.[10]

[10] *Journal of Chemical Information and Modelling* 2013, 53, 11, 2884-2895.

Graph isomorphism algorithms are widely used for the atom mapping job.

In the example below, identification of the reaction center and the atom mapping does not seem too difficult, however, this task becomes very difficult when a generic reaction is presented in a publication (using R groups and tables for reagents) or when the reaction is incomplete.

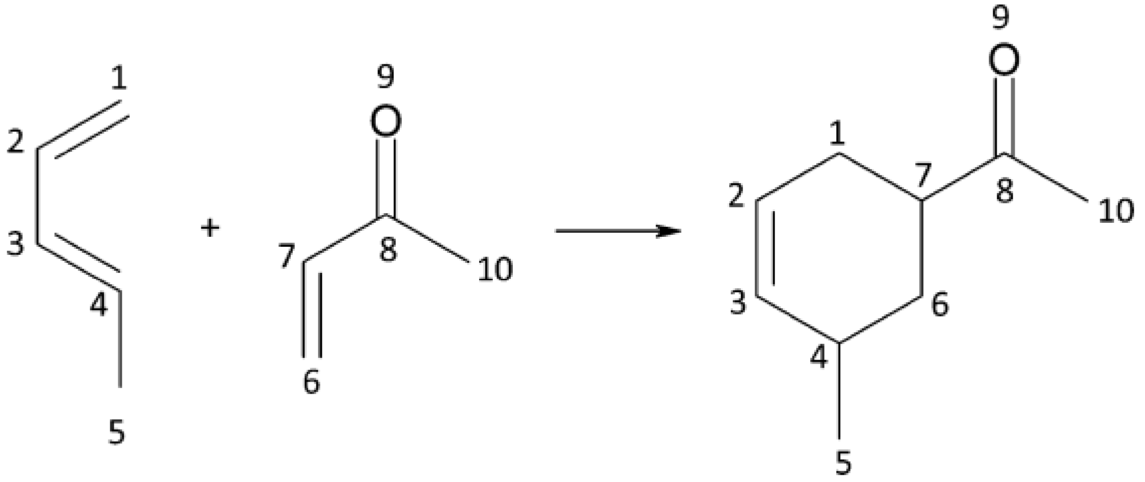


Figure 1. Atoms 1–4, 6, and 7 are in the reaction center; bonds 4–5, 7–8, 8–9, and 8–10 are unchanged.

Figure 5. This is a very simplistic example of the reaction mapping process. A Diels-Alder reaction.

Another example of more complexity can be seen below:

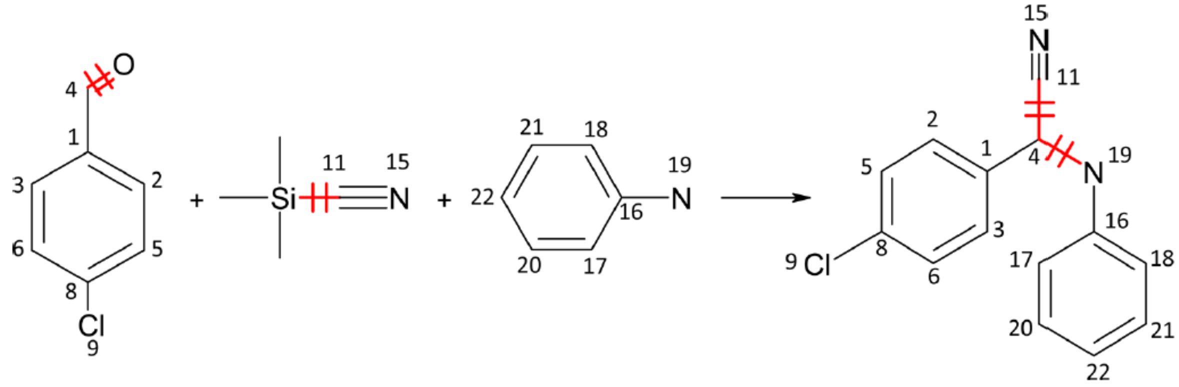


Figure 6. In this example the reaction is not adjust (same number of atoms at both sides of the arrow). However, a good reaction mapping algorithm must be able to get it right.

A computer must understand that the Si and the O atoms are no longer present in the product - a computer does not know reaction mechanisms like we do!

Once the reaction center is identified, an algorithm generates a **classification code**. This code represents only the changes produced at the reaction center. Atoms and bonds not involved in the reaction center are not used for a reaction classification process (in computers-world).

4. CASD SYSTEMS FOR RETROSYNTHETIC ANALYSIS (CHEMPLANNER & ICSYNTH)

List of available CASD systems:

- LHASA[11]
- SYNCHEM
- COMPASS
- WODCA
- SST
- KOSP
- HORACE
- SECS
- FLAMINGOES
- EROS
- SYNGEN[12]
- CHIRON
- **ARChem / Chem Planner**
- **ICSYNTH**

Not CASD but synthesis planners:

- Reaxys Synthesis Planner
- SciFinder, SciPlanner
- Schematica

a) ICSYNTH (by InfoChem)[13]

How does it work?

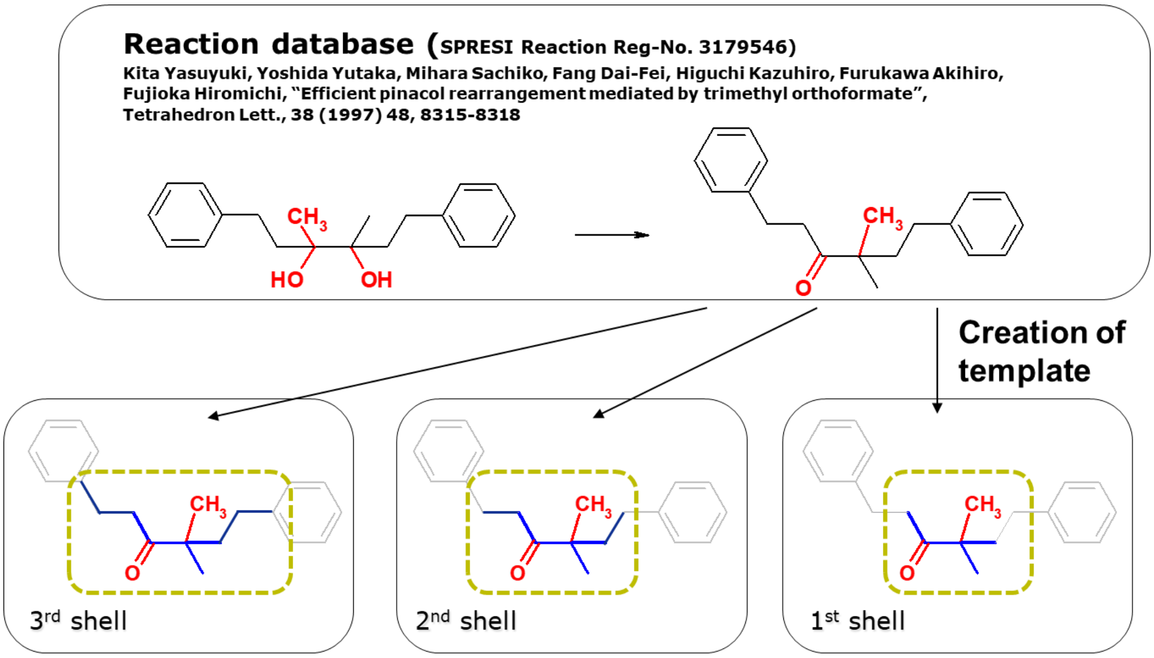
Preprocessing of a reaction database identify reactions and generate *templates* for both the products and the starting materials. These templates are generated in up to 3 different levels (shells).

Once the templates are generated, a transform describes the differences between the starting materials and the products.

[11] http://cheminf.cmbi.ru.nl/cheminf/lhasa/doc/lhasa191.pdf

[12] https://www.researchgate.net/profile/James_Hendrickson/publication/239588231_Synthesis_design_logic_and_the_SYNGEN_synthesis_generation_program/links/00b495287b39fa7742000000/Synthesis-design-logic-and-the-SYNGEN-synthesis-generation-program.pdf

[13] https://www.infochem.de/

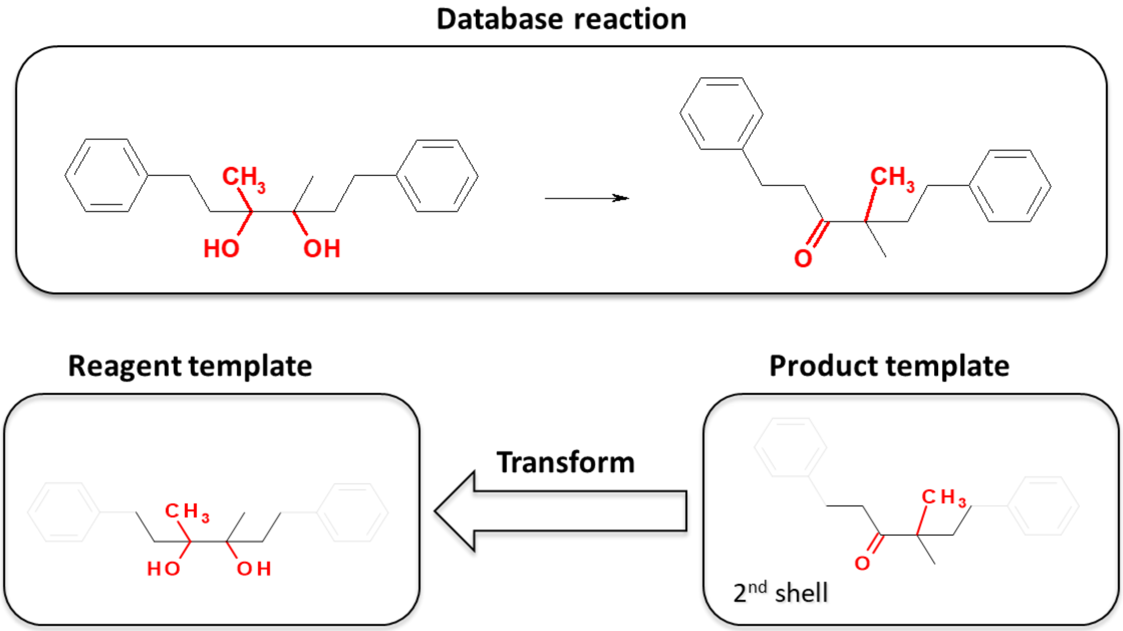


Reaction database (SPRESI Reaction Reg-No. 3179546)

Kita Yasuyuki, Yoshida Yutaka, Mihara Sachiko, Fang Dai-Fei, Higuchi Kazuhiro, Furukawa Akihiro, Fujioka Hiromichi, "Efficient pinacol rearrangement mediated by trimethyl orthoformate", Tetrahedron Lett., 38 (1997) 48, 8315-8318

- Templates are created from both products and precursors

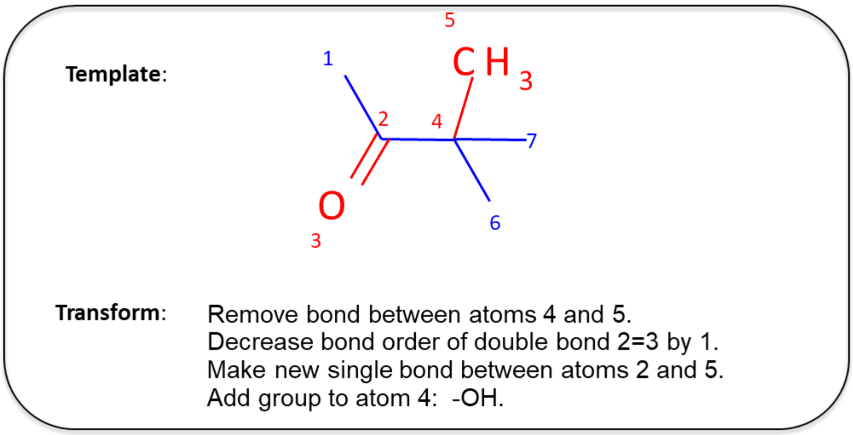
Figure 7.Creation of templates from a reaction database.



- A chemical transformation rule (transform) is derived

Figure 8. A transform describes the changes between the starting materials and the products.

The transform is then written and saved in a transform library.



Transform: Remove bond between atoms 4 and 5.
Decrease bond order of double bond 2=3 by 1.
Make new single bond between atoms 2 and 5.
Add group to atom 4: -OH.

- Reactions of the same type are grouped together and a chemical transformation rule is derived (transform)
- The transform describes the structural difference between ’before’ and ’after’ snapshots, but not the mechanistic process

Figure 9. A transform describes atoms and bonds changes. No mechanistic knowledge is involved in the process.

Finally, when a query is sent to IC*SYNTH*, the system compares the query structure template with those saved in the database, applies the corresponding transform and generates suggested precursors for the target (Figure 10).

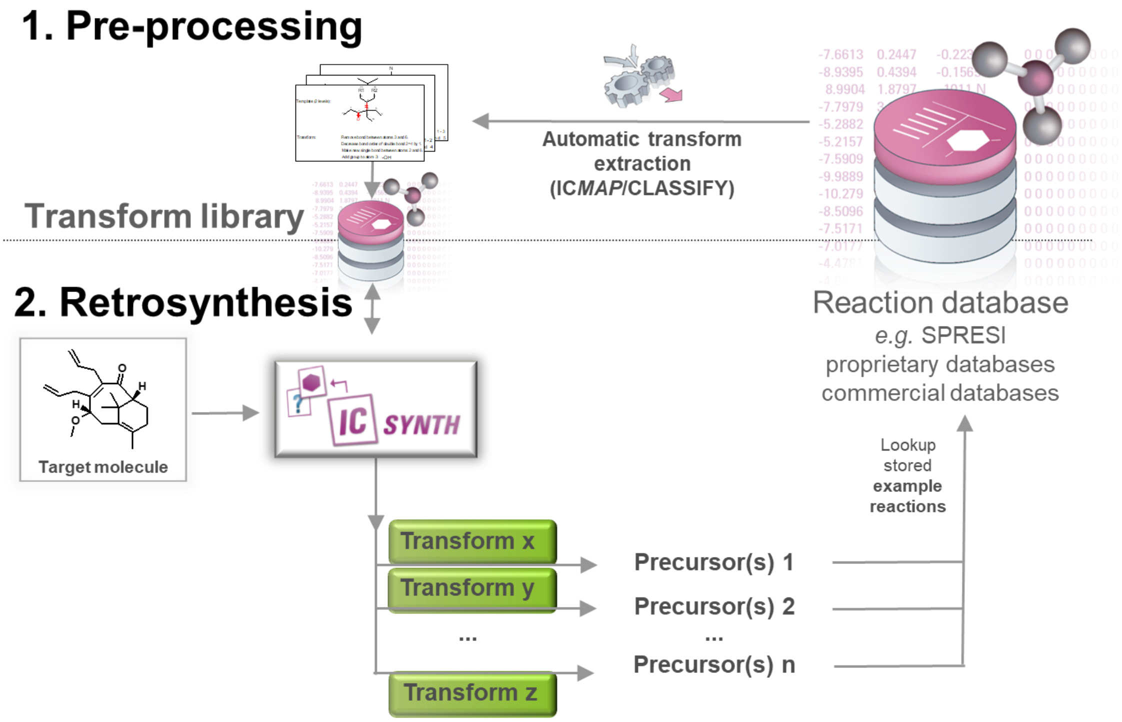


Figure 10. ICSynth in action. From input of the query (Target molecule) to the suggestion of precursors with the corresponding examples in the literature.

b) ChemPlanner (Wiley)

Originally developed by Simbiosys and named ARChem. It was a natural evolution of the original LHASA system developed by Corey.

In essence, it shares many common features with IC*SYNTH*, the main difference being the transform library generation. ChemPlanner applies the same principles that Corey described back in 1969, but it does it very well.

Nowadays *ChemPlanner* is one of the most reliable CASD systems out there.

Latest news indicate that ChemPlanner has been acquired by CAS and it will be integrated in SciFinder (at an additional cost!!).

As far as this text material was written, the original web page www.chemplanner.com does not exist any longer and it redirects to: https://www.cas.org/products/scifinder/retrosynthesis-planning

- Some examples at: https://scifinder-n.cas.org/

ChemPlanner_Technical_Notes.pdf   ChemPlannerStarkWhitePaper_WEB.pdf

c) When and how to use them?

Of course, these systems can be used whenever and for whatever someone wants to use them (they are very expensive, so it's good to use them). Having said that, I will "only" use these systems for unknown molecules with no close analogues described in the literature, or to come up with "out of the box" new synthetic routes.

Which system is better?

Spite off the huge similarity in the operating algorithm of the different CASD systems, there are some details that, in certain cases, sets them apart:

1) Systems based on pure graph theory with little chemistry knowledge (rules).
2) Systems based on heuristics and rules.

The first ones are slightly better for “out of the box” solutions. However, they tend to give more “noise” than those based on heuristics.

The second ones are more reliable and robust. Very useful for close analogues to molecules already reported. In addition, quite often these systems provide better starting materials selection.

5. OTHER CASD SYSTEMS

a) IC*FRP* and Muse Invent

Computer Aided Drug Design (CADD) + Synthetic Chemistry

Two systems to virtually create small molecules of potential biological interest.

IC*FRP*[14] works using IC*SYNTH* algorithms but in reverse mode, applying reactions to a query molecule (starting material).

Muse Invent[15] has a set of up to 80 reactions that are applied to a given query structure (acting as starting material) and then generates molecules with biological interest (filtering the possible products through basic med chem rules).

b) AI in synthesis design[16]

Latest advances in the field of retrosynthetic analysis comes from the Artificial Intelligence algorithms (specially from the Machine Learning applications).

For this type of analysis, the Network of Organic Chemistry (NOC) is widely used.[17]

As an example of how is used, see example below (Figure 11).

[14] https://www.infochem.de/synthesis/ic-frp

[15] https://www.certara.com/2014/10/15/synthetic-chemistry-cadd-muse-invent/?

[16] a) Baskin, I. I., Madzhidov, T. I., Antipin, I. S., & Varnek, A. A. (2017). Artificial intelligence in synthetic chemistry: achievements and prospects. Russian Chemical Reviews, 86(11), 1127–1156. https://doi.org/10.1070/rcr4746. b) de Almeida, A. F., Moreira, R., & Rodrigues, T. (2019). Synthetic organic chemistry driven by artificial intelligence. Nature Reviews Chemistry, 3(10), 589–604. https://doi.org/10.1038/s41570-019-0124-0

[17] Grzybowski, B. A., Bishop, K. J. M., Kowalczyk, B., & Wilmer, C. E. (2009). The "wired" universe of organic chemistry. *Nature Chemistry*, 1(1), 31–36. https://doi.org/10.1038/nchem.136

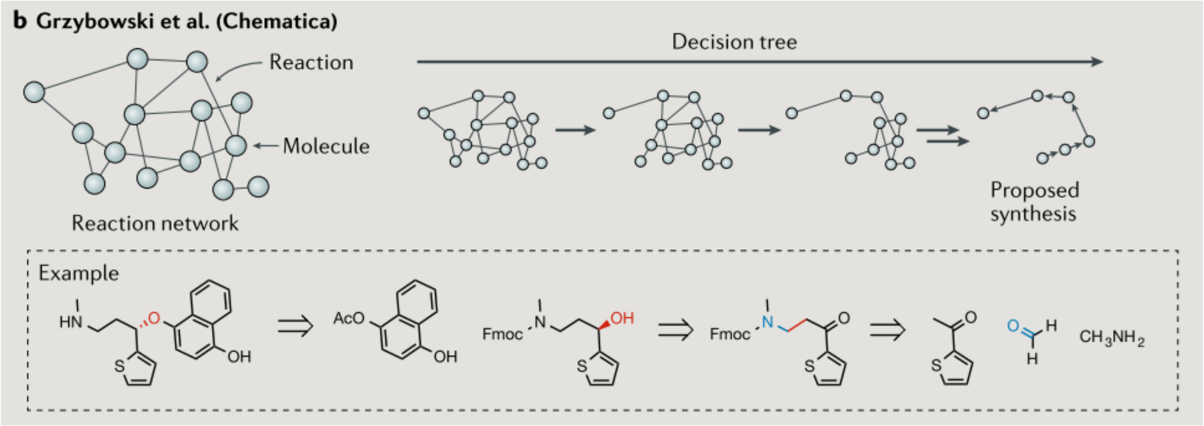


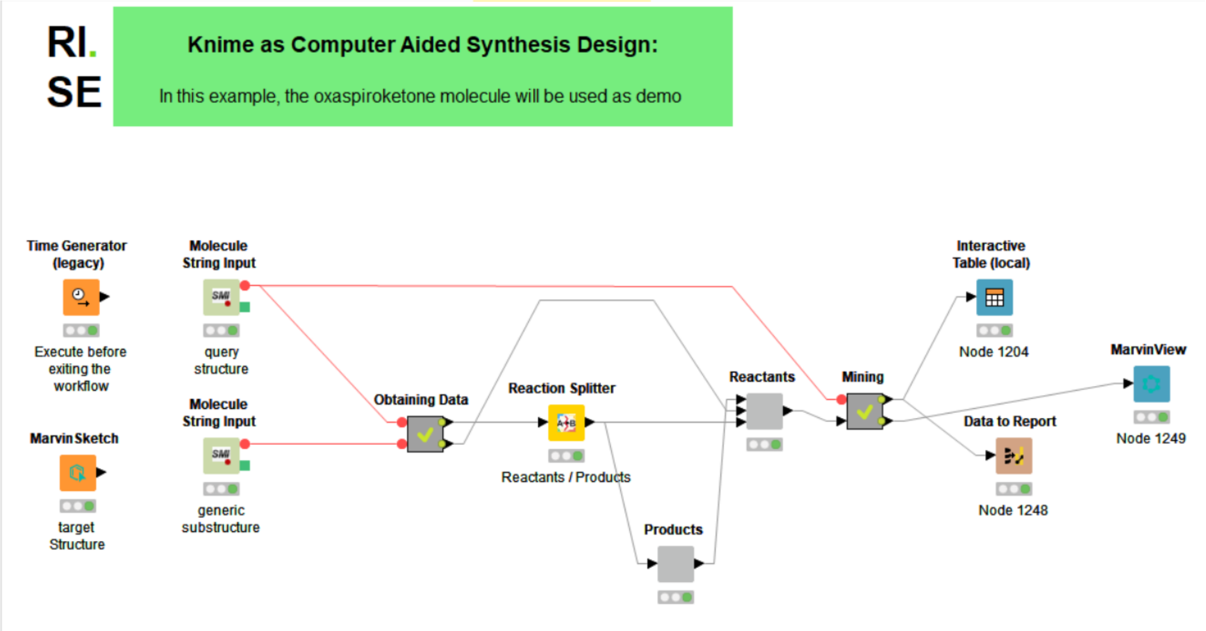
Figure 11. Application of the NOC to the synthesis of a target molecule.

6. OTHER TOOLS FOR SYNTHESIS (HOMEMADE)

Despite the development and existence of CASD systems, most of them (almost all) are commercial and require of an important investment from private companies and/or research organizations.

Knowing how CASD systems work, it is possible to use that knowledge to get the right information (and ideas) from the reaction databases available today ().

Think that the suggestions from a CASD system are based on reactions found in the databases.



Knime as Computer Aided Synthesis Design:

In this example, the oxaspiroketone molecule will be used as demo

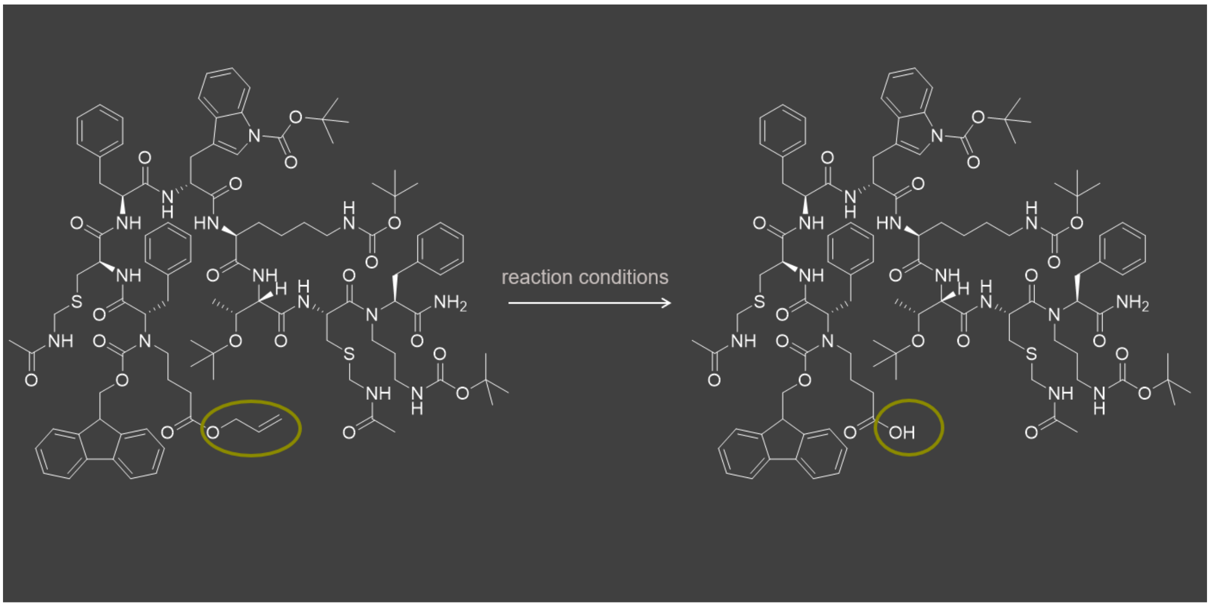
Figure 12. Example of a CASD system developed using KNIME and Reaxys (API).

More common search engines such as SciFinder and Reaxys are an excellent help for synthetic chemists, however, they lack flexibility and they are not a big help in solving particular (and specific) synthetic problems such as:

a) Reaction Conditions Search
b) Synthetic Methodology Search
c) Due Diligence (combining searches) -> Med Chem specific
d) Batch query searches

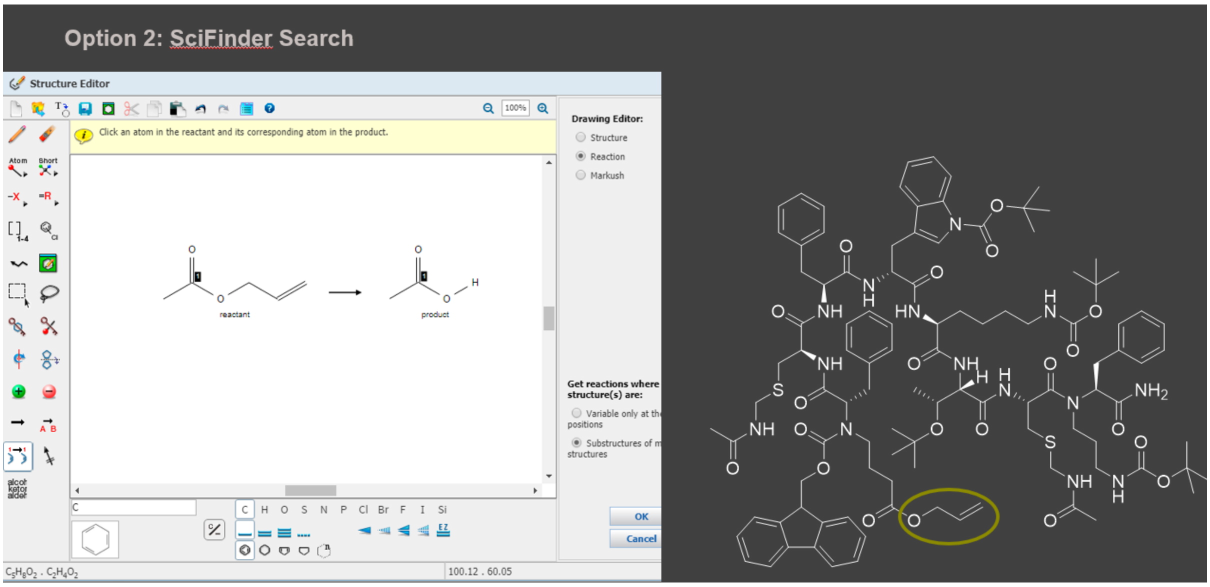
a) Reaction Conditions Search

Finding new reaction conditions for the hydrolysis of an ester in a complex molecule can be a tedious problem.



The first option will be to search in a database for the hydrolysis of a more generic molecule.

However, a generic search will give us ca 700k hits, which it's not very practical.



Option 2: SciFinder Search

Structure Editor

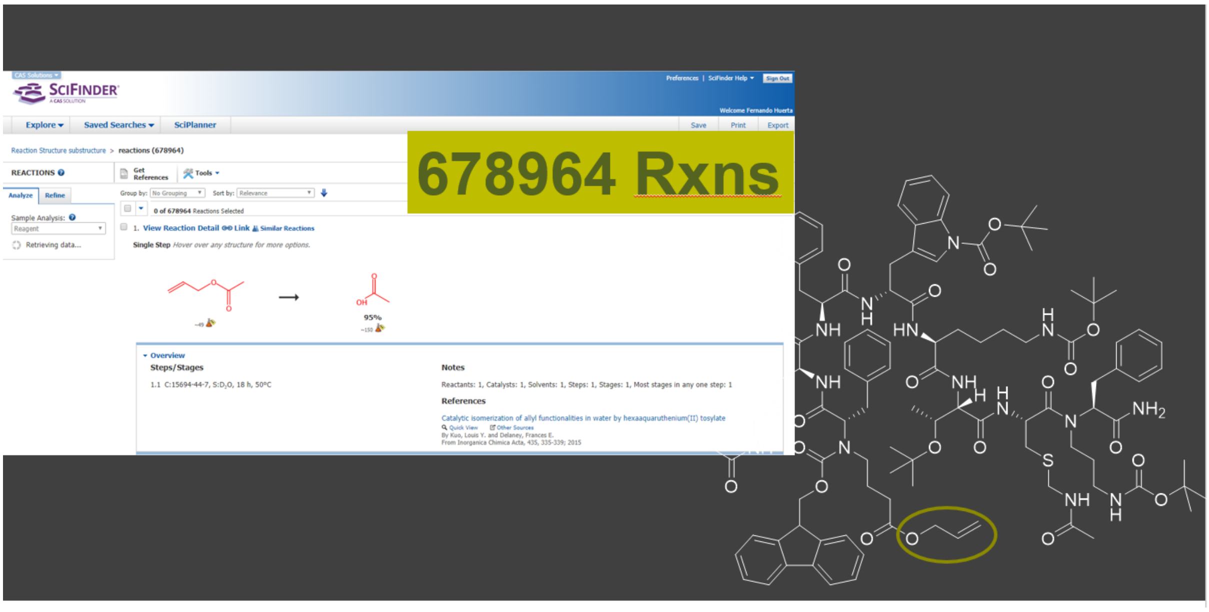
Click an atom in the reactant and its corresponding atom in the product.

Drawing Editor:
○ Structure
● Reaction
○ Markush

Get reactions where structure(s) are:
○ Variable only at th positions
● Substructures of m structures

$C_5H_8O_2 . C_2H_4O_2$

100.12 . 60.05



Reaction Structure substructure > reactions (678964)

REACTIONS

Group by: No Grouping
Sort by: Relevance

0 of 678964 Reactions Selected

1. View Reaction Detail Link Similar Reactions

Single Step *Hover over any structure for more options.*

Sample Analysis: Reagent

Retrieving data...

Overview

Steps/Stages

1.1 C:15694-44-7, S:D_2O, 18 h, 50°C

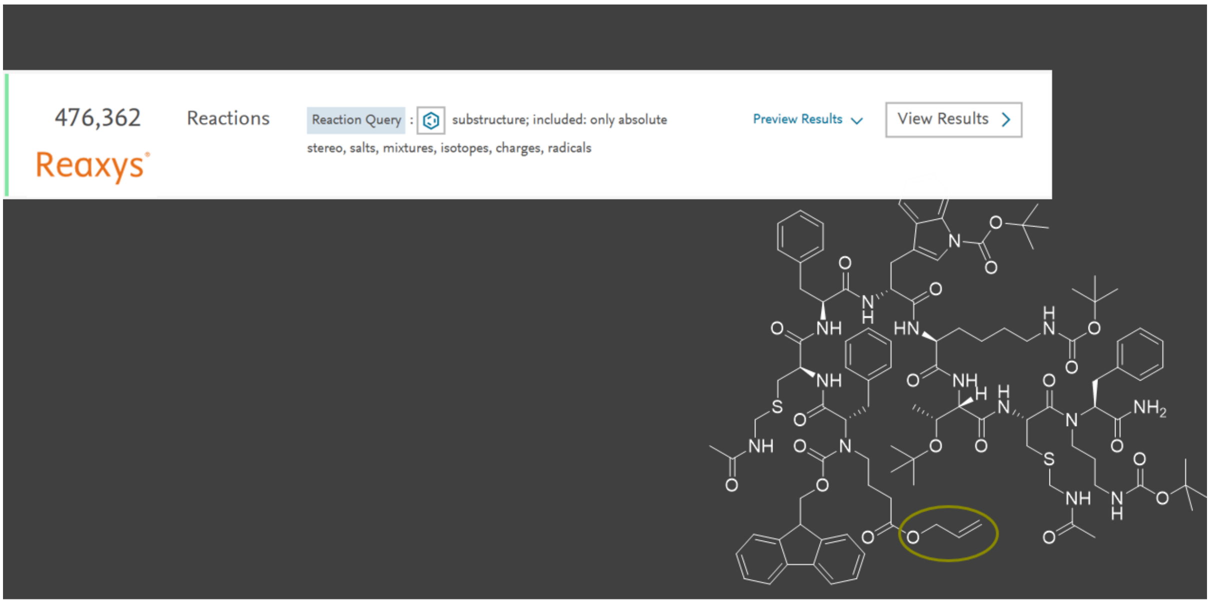
Notes

Reactants: 1, Catalysts: 1, Solvents: 1, Steps: 1, Stages: 1, Most stages in any one step: 1

References

Catalytic isomerization of allyl functionalities in water by hexaaquaruthenium(II) tosylate
By Kuo, Louis Y. and Delaney, Frances E.
From Inorganica Chimica Acta, 435, 335-339; 2015

The alternative is to use Reaxys instead of SciFinder, but not much is gained, ca 500k hits are found in Reaxys.



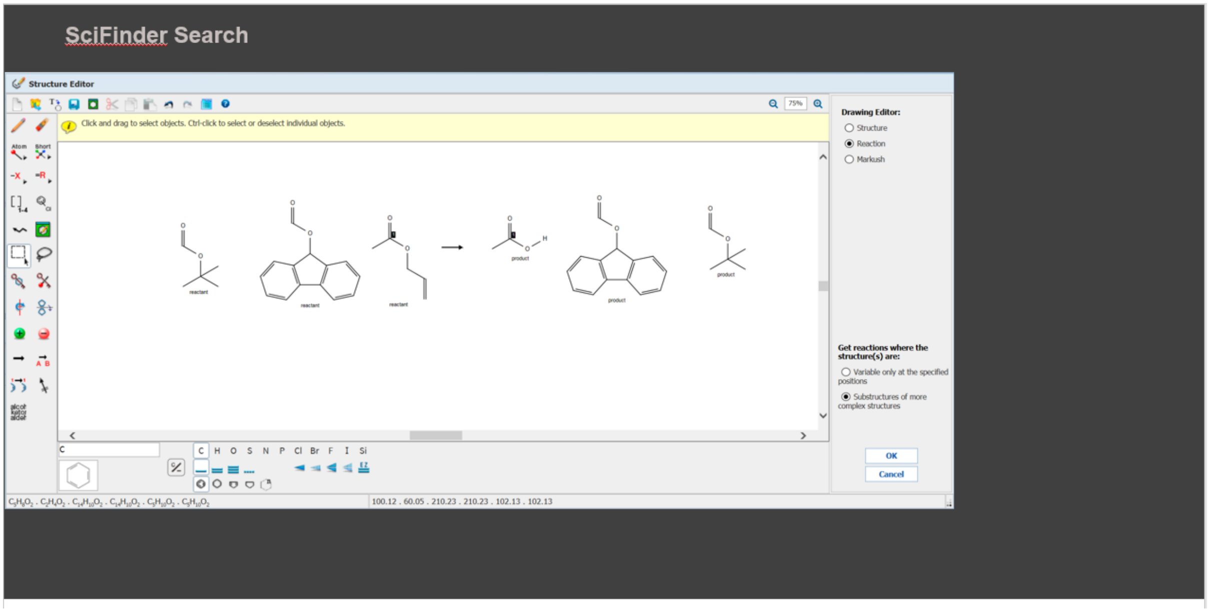
476,362 Reactions

Reaction Query : substructure; included: only absolute stereo, salts, mixtures, isotopes, charges, radicals

Preview Results

View Results

The best alternative is to be more specific (but without having to draw the full molecule).



SciFinder Search

Structure Editor

Click and drag to select objects. Ctrl-click to select or deselect individual objects.

Drawing Editor:
○ Structure
● Reaction
○ Markush

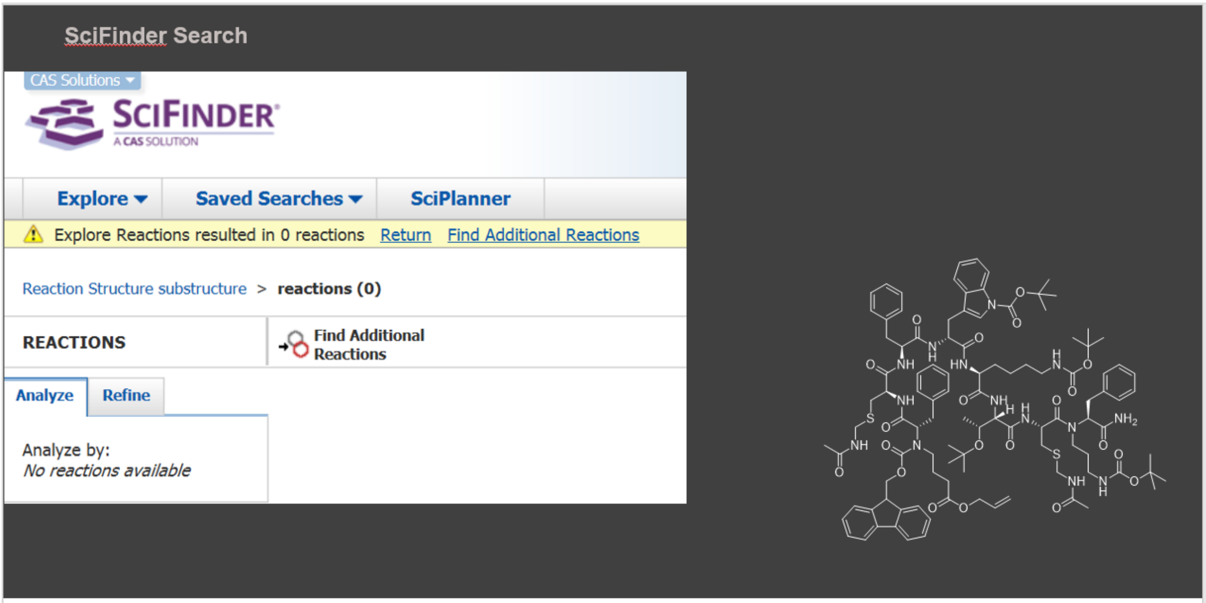
Get reactions where the structure(s) are:
○ Variable only at the specified positions
● Substructures of more complex structures

OK
Cancel

$C_5H_8O_2$. $C_2H_4O_2$. $C_{14}H_{10}O_2$. $C_{14}H_{10}O_2$. $C_5H_{10}O_2$. $C_5H_{10}O_2$

100.12 . 60.05 . 210.23 . 210.23 . 102.13 . 102.13

This option is not really good either.



SciFinder Search

Explore Reactions resulted in 0 reactions Return Find Additional Reactions

Reaction Structure substructure > reactions (0)

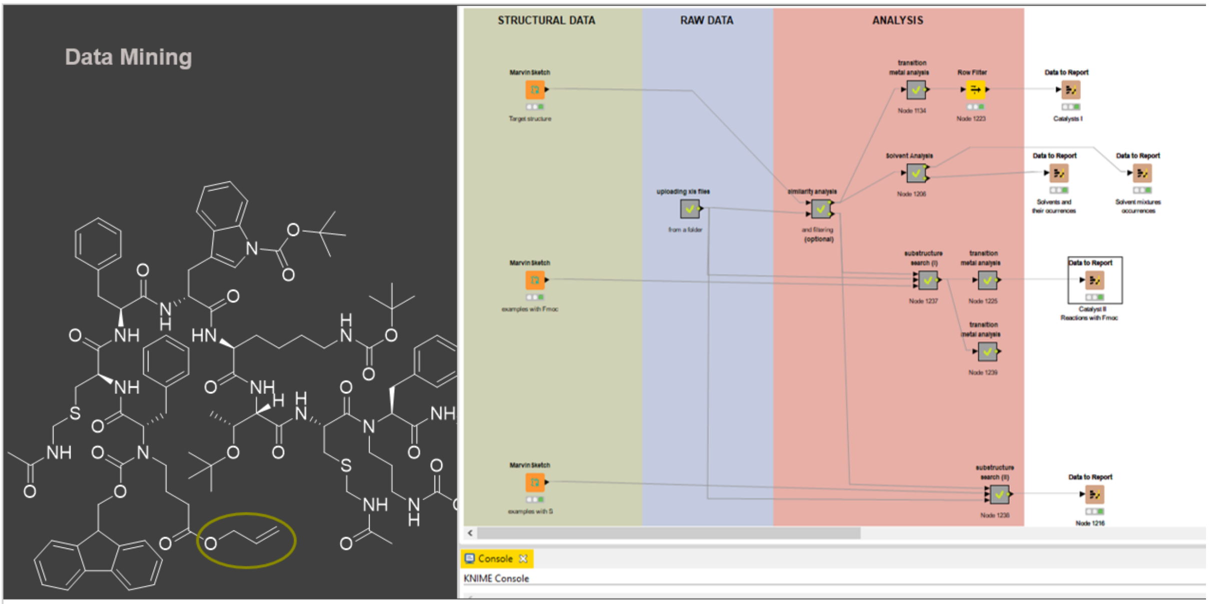
REACTIONS

Find Additional Reactions

Analyze by:
No reactions available

No hits are found when we go specific.

So, the good news is the existence of open source software that allow us to handle these 700.000 hits and find exactly what we want.



Data Mining

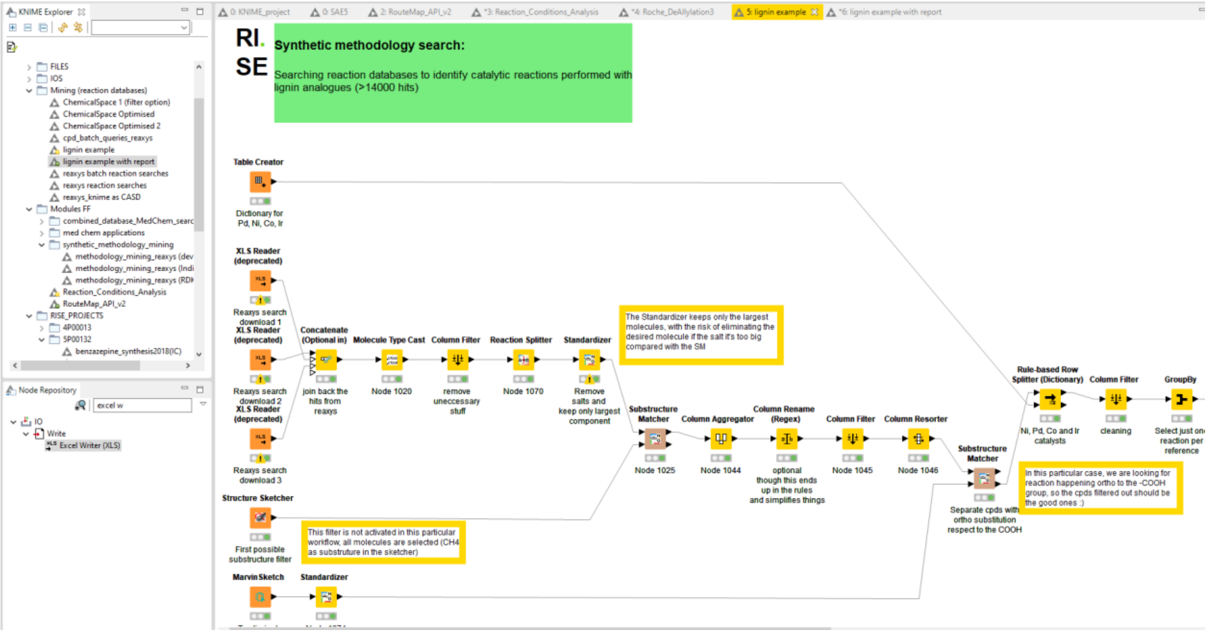
Using an analytical platform such as KNIME, and direct access to Reaxys (or any reaction database) via an API (application program interface), it's possible to analyze all chemical information in the database and find exactly what is needed.[18]

https://www.knime.com/downloads/download-knime

For a KNIME video tutorial see: https://www.youtube.com/watch?v=W2WTPzwIqv8

b) Synthetic Methodology Search

Example1



Synthetic methodology search:

Searching reaction databases to identify catalytic reactions performed with lignin analogues (>14000 hits)

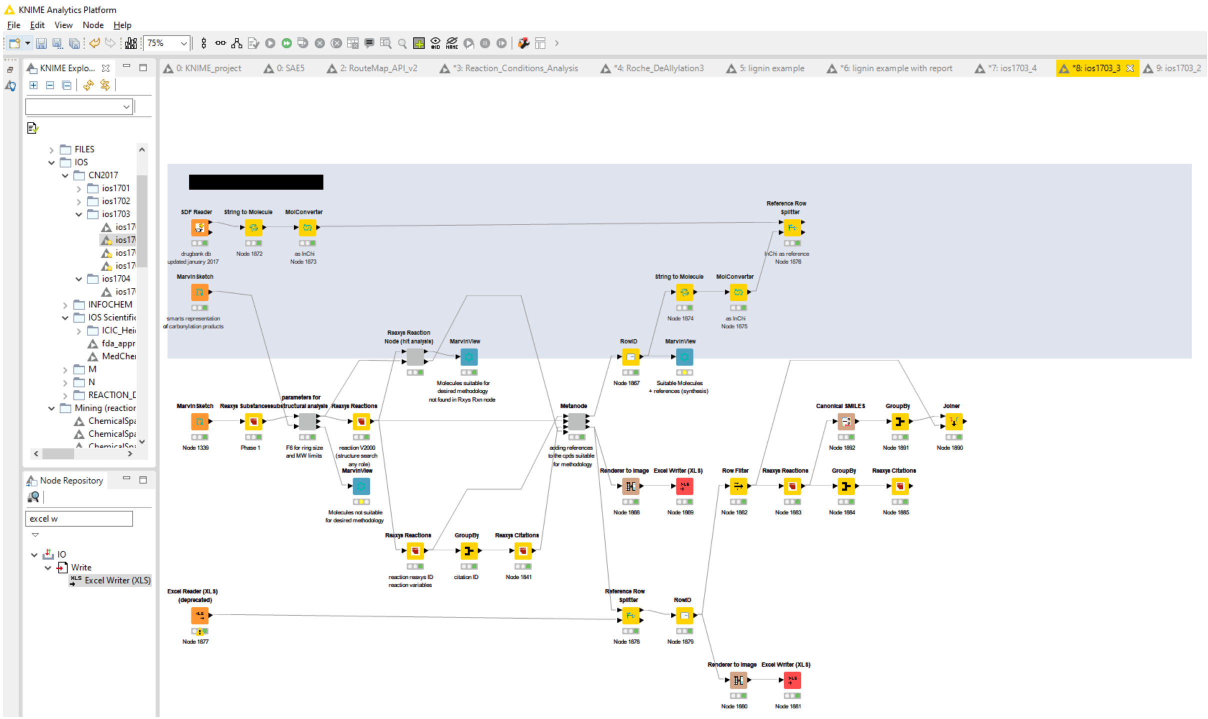
The Standardizer keeps only the largest molecules, with the risk of eliminating the desired molecule if the salt it's too big compared with the SM

This filter is not activated in this particular workflow, all molecules are selected (CH4 as substruture in the sketcher)

In this particular case, we are looking for reaction happening ortho to the -COOH group, so the cpds filtered out should be the good ones :)

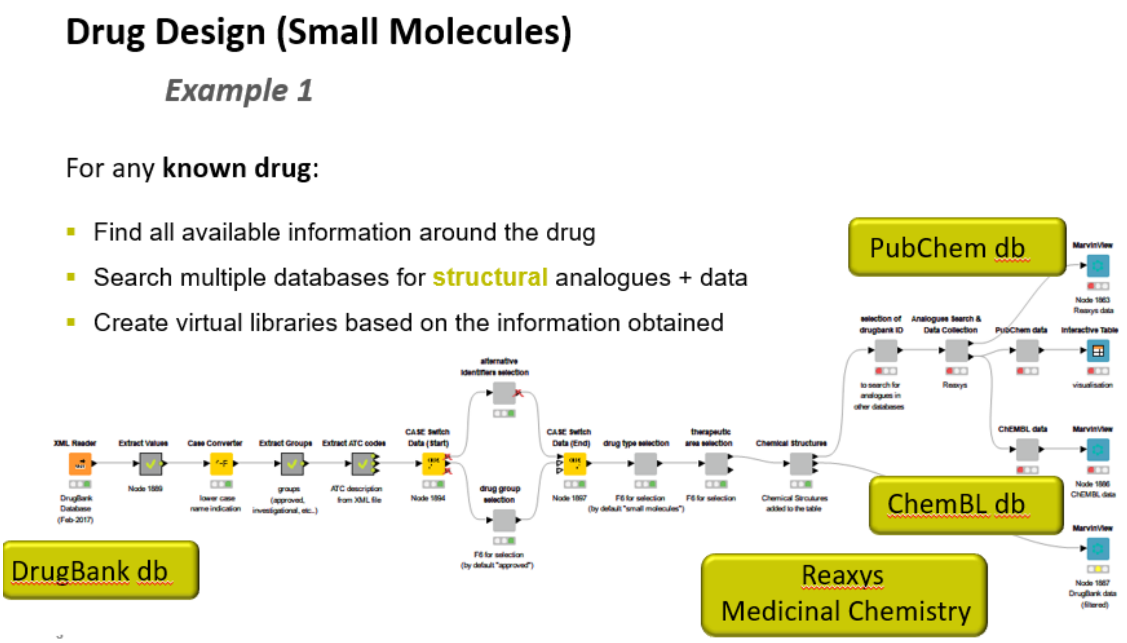
Example2. Search for all molecules in phase 1 (clinical trials) that could be synthesized via a carbonylation reaction.

[18] a) For students with programming knowledge, scripting languages such as *R* or *Python* would allow for the same type of analysis. b) Other analytical platforms also possible, see; https://alternativeto.net/software/knime/.



c) Due Diligence (combine searches)

Not really part of this course, but still a good use of cheminformatic tools.



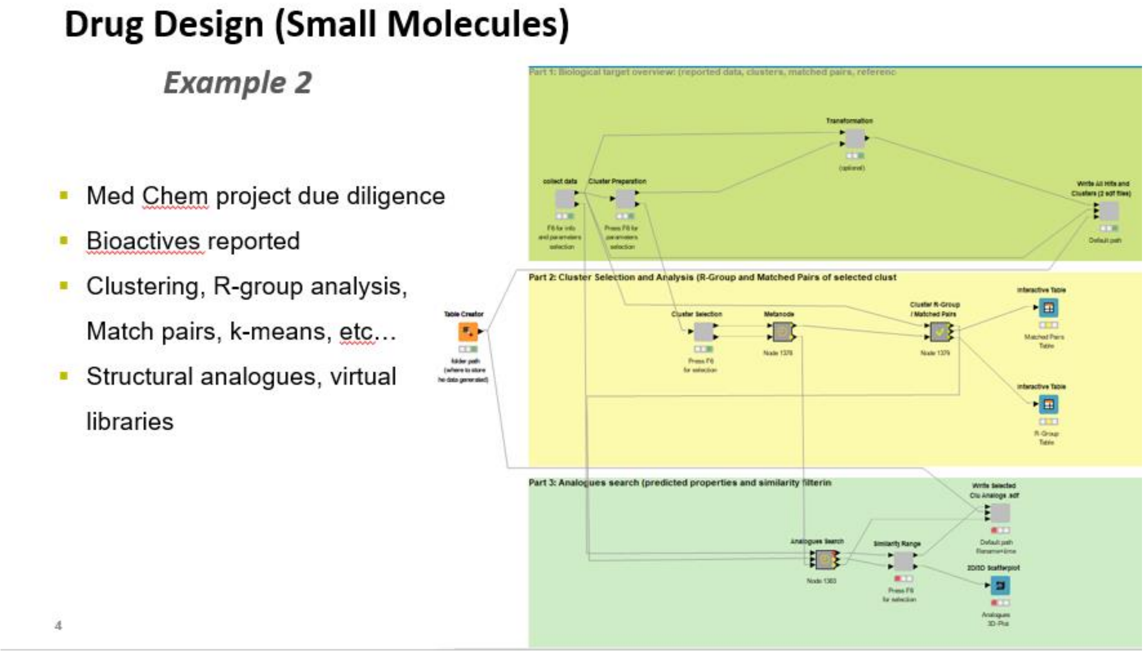
Drug Design (Small Molecules)

Example 1

For any **known drug**:

- Find all available information around the drug
- Search multiple databases for **structural** analogues + data
- Create virtual libraries based on the information obtained

Figure 13. Combined search to find all active (molecular entities) in a target receptor.



Drug Design (Small Molecules)

Example 2

- Med Chem project due diligence
- Bioactives reported
- Clustering, R-group analysis, Match pairs, k-means, etc…
- Structural analogues, virtual libraries

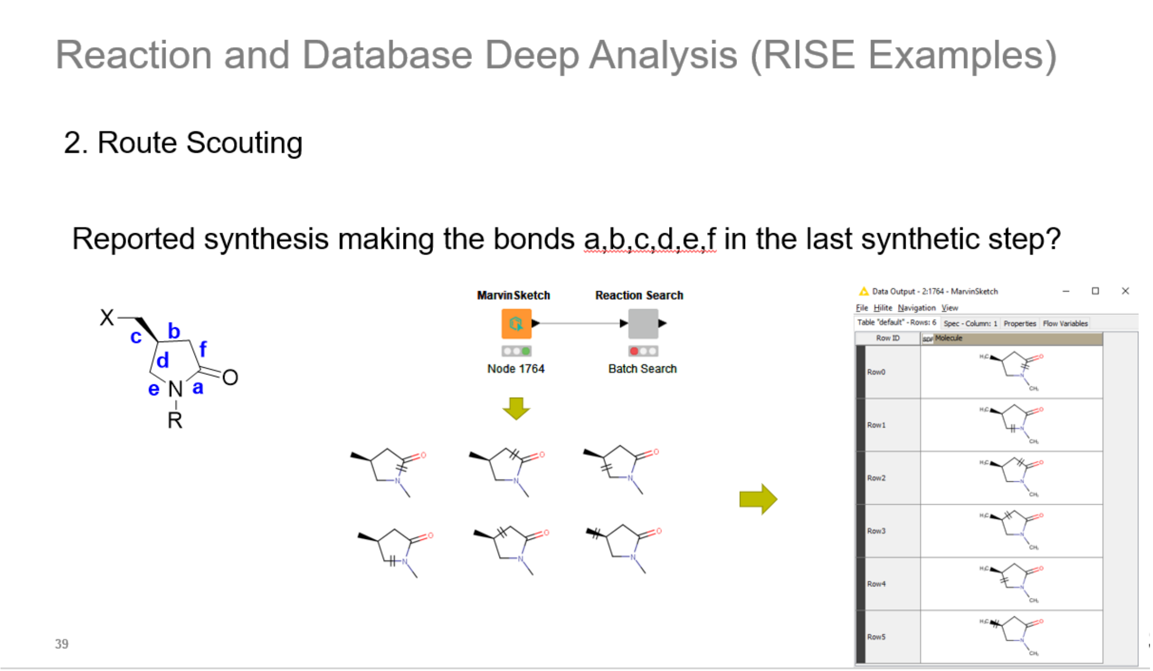
4

Part 1: Biological target overview: (reported data, clusters, matched pairs, referenc
Transformation
(optional)
collect data
F6 for info and parameters selection
Cluster Preparation
Press F6 for parameters selection
Write All Hits and Clusters (2 sdf files)
Default path
Part 2: Cluster Selection and Analysis (R-Group and Matched Pairs of selected clust
Table Creator
folder path (where to store he data generated)
Cluster Selection
Press F6 for selection
Metanode
Node 1378
Cluster R-Group / Matched Pairs
Node 1379
Interactive Table
Matched Pairs Table
Interactive Table
R-Group Table
Part 3: Analogues search (predicted properties and similarity filterin
Write Selected Clu Analogs .sdf
Default path Rename+time
Analogues Search
Node 1383
Similarity Range
Press F6 for selection
2D/3D Scatterplot
Analogues 3D-Plot

Figure 14. All information collected in Figure 13 can be used to generate a project overview and finding new virtual molecules with the desired med chem properties.

d) Batch Query Search

Searching simultaneously for all 6 disconnections in the slide below is not an easy task using any of the commercially available databases.



Reaction and Database Deep Analysis (RISE Examples)

2. Route Scouting

Reported synthesis making the bonds a,b,c,d,e,f in the last synthetic step?

39

Using KNIME the search is all done at once and the "unwanted" results are filtered, saving a lot of time for doing chemistry in the lab.

7. REFERENCES (EXTRA READING)

Below a list of more bibliographic references related to the course.

a) Coley, C. W., Green, W. H., & Jensen, K. F. (2018). Machine Learning in Computer-Aided Synthesis Planning [Research-article]. *Accounts of Chemical Research*, *51*(5), 1281–1289. https://doi.org/10.1021/acs.accounts.8b00087

b) Graulich, N., Hopf, H., & Schreiner, P. R. (2010). Heuristic thinking makes a chemist smart. *Chemical Society Reviews*, *39*(5), 1503–1512. https://doi.org/10.1039/b911536f

c) Roughley, S. D., & Jordan, A. M. (2011). The medicinal chemist's toolbox: An analysis of reactions used in the pursuit of drug candidates. *Journal of Medicinal Chemistry*, *54*(10), 3451–3479. https://doi.org/10.1021/jm200187y

d) Szymkuć, S., Gajewska, E. P., Klucznik, T., Molga, K., Dittwald, P., Startek, M., … Grzybowski, B. A. (2016). Computer-Assisted Synthetic Planning: The End of the Beginning. *Angewandte Chemie - International Edition*, Vol. 55. https://doi.org/10.1002/anie.201506101

e) Avramova, S., Kochev, N., & Angelov, P. (2018). RetroTransformDB: A Dataset of Generic Transforms for Retrosynthetic Analysis. *Data*, *3*(2), 14. https://doi.org/10.3390/data3020014

f) Wei, J. N., Duvenaud, D., & Aspuru-Guzik, A. (2016). Neural networks for the prediction of organic chemistry reactions. *ACS Central Science*, *2*(10), 725–732. https://doi.org/10.1021/acscentsci.6b00219

g) Segler, M. H. S., & Waller, M. P. (2017). Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chemistry - A European Journal*, *23*(25), 5966–5971. https://doi.org/10.1002/chem.201605499

h) Feng, F., Lai, L., & Pei, J. (2018). Computational chemical synthesis analysis and pathway design. *Frontiers in Chemistry*, *6*(JUN). https://doi.org/10.3389/fchem.2018.00199

i) Boda, K., Seidel, T., & Gasteiger, J. (2007). Structure and reaction based evaluation of synthetic accessibility. *Journal of Computer-Aided Molecular Design*, *21*(6), 311–325. https://doi.org/10.1007/s10822-006-9099-2

j) Bøgevig, A., Federsel, H.-J., Huerta, F., Hutchings, M. G., Kraut, H., Langer, T., … Saller, H. (2015). Route Design in the 21st Century: The IC SYNTH Software Tool as an Idea Generator for Synthesis Prediction. *Organic Process Research & Development*, 19(2), 357–368. https://doi.org/10.1021/op500373e

k) Gothard, C. M., Soh, S., Gothard, N. A., Kowalczyk, B., Wei, Y., Baytekin, B., & Grzybowski, B. A. (2012). Rewiring chemistry: algorithmic discovery and experimental validation of one-pot reactions in the network of organic chemistry. *Angewandte Chemie (International Ed. in English)*, *51*(32), 7922–7927. https://doi.org/10.1002/anie.201202155

l) Szymkuć, S., Gajewska, E. P., Klucznik, T., Molga, K., Dittwald, P., Startek, M., … Grzybowski, B. A. (2016). Computer-Assisted Synthetic Planning: The End of the Beginning. *Angewandte Chemie - International Edition*, Vol. 55. https://doi.org/10.1002/anie.201506101