Computer-Assisted Structure Elucidation

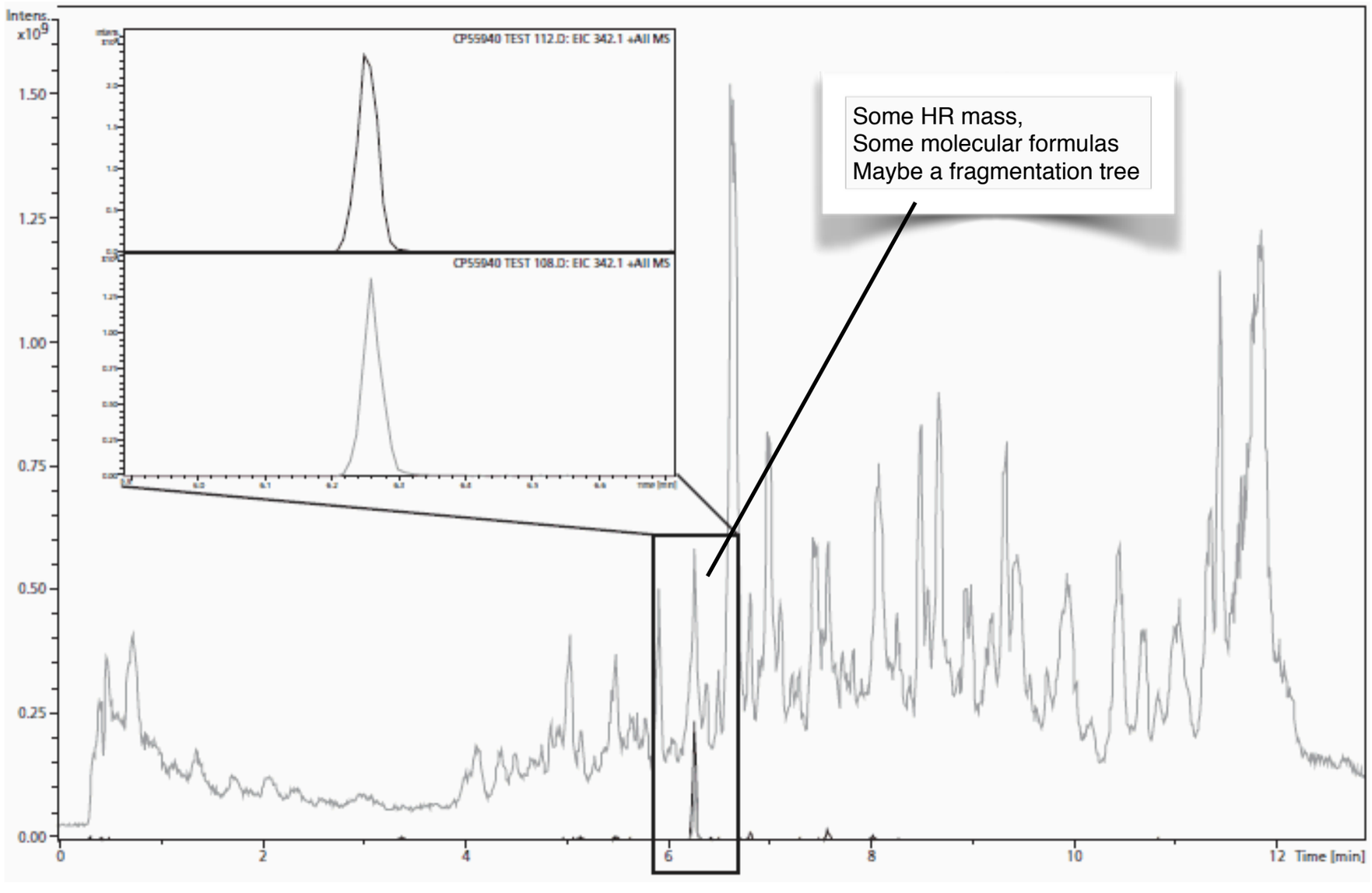
Christoph Steinbeck

In a typical metabolome measurement, less than 40% of the features can be assigned to known compounds.

Oliver Fiehn, UC Davis, USA
Internal communication

Setting the scene

(i.e., what are we actually talking about?)



I have a molecular formula from HR MS.

My database has a hit (or two, or three).

Problem solved.

Or is it?

There are **known knowns**; there are things we know we know.

We also know there are **known unknowns**; that is to say, we know there are some things we do not know.

But there are also **unknown unknowns** – the ones we don't know we don't know.

—United States Secretary of Defense,

Donald Rumsfeld

Levels of Confidence

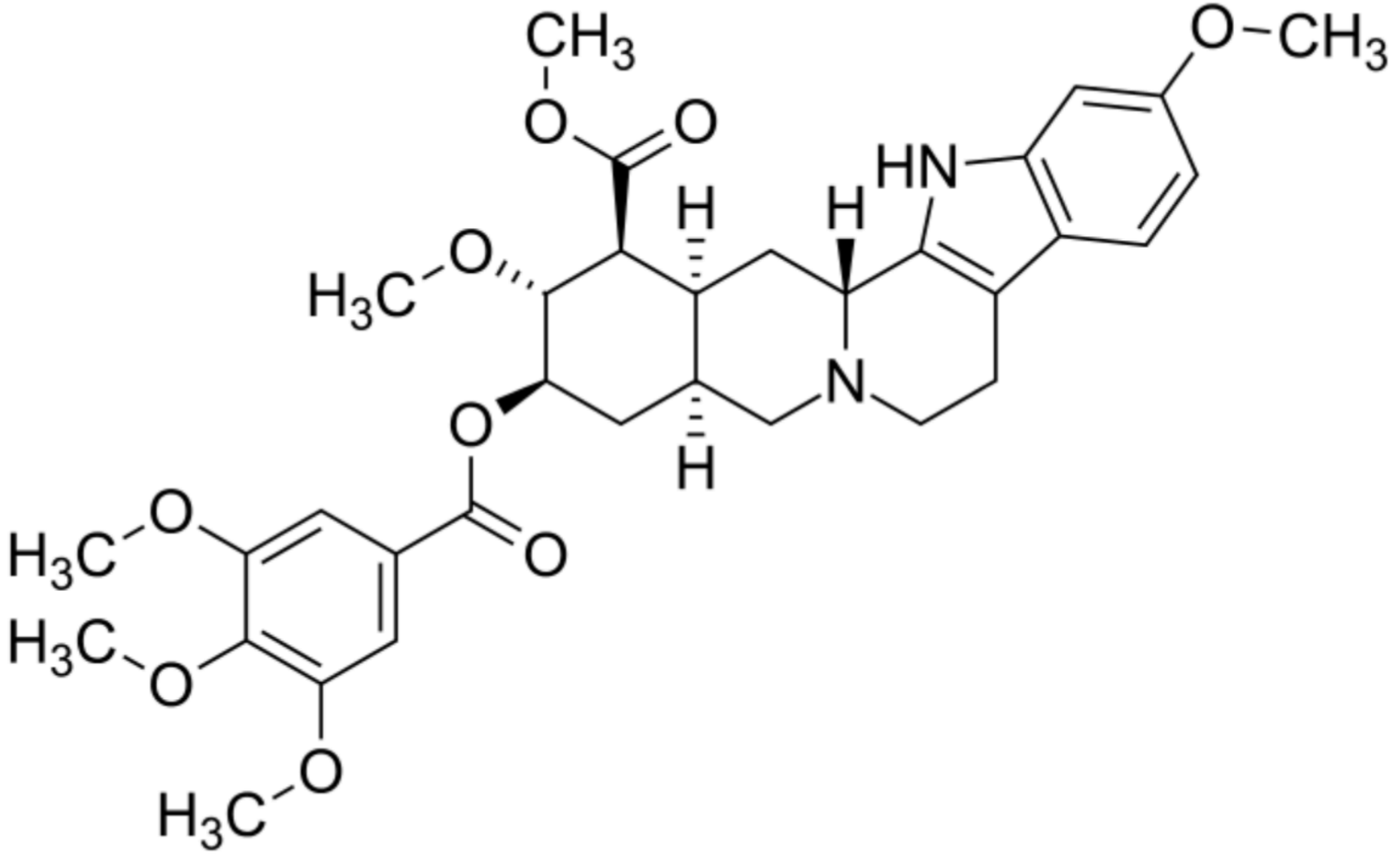
Table 1

Summary of levels of confidence in metabolite 'identification', as defined by the Chemical Analysis Working Group of the Metabolomics Standards Initiative [4]

Level of confidence	Description	Requisite analytical data
Level 1	'Identified metabolites'	Two orthogonal analytical techniques applied to the analysis of both the metabolite of interest and to a chemical reference standard of suspected structural equivalence, with all analyses performed under identical analytical conditions within the same laboratory Examples of appropriate orthogonal data: accurate mass via MS with retention time; accurate mass MS and fragmentation data or isotopic pattern; 2D NMR spectra; full ^{1}H and/or ^{13}C NMR spectra and so on
Level 2	'Putatively annotated compounds'	As for levels 3 and 4, including spectral (NMR and/or MS) similarity with public or commercial libraries
Level 3	'Putatively characterised compound classes'	As for level 4, plus spectral and/or physicochemical properties consistent with a particular class of organic compounds
Level 4	'Unknown'	A discernible spectral signal (NMR, MS or other) that can be reproducibly detected and quantified

Reproduced from: Viant MR, Kurland IJ, Jones MR, Dunn WB (2017) How close are we to complete annotation of metabolomes? Current Opinion in Chemical Biology 36:64–69. doi: 10.1016/j.cbpa.2017.01.001

The More Complicated Cases



So to be clear:

For everything that follows, we need a **pure**, **isolated** compound in **sufficient amounts**.

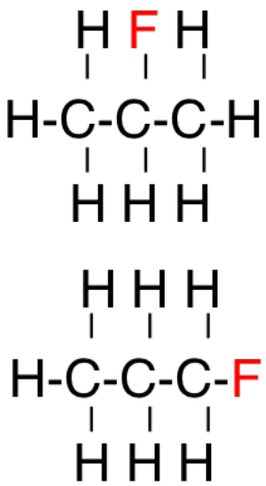
Isomers

Isomers

share one molecular formula

Isomers

share one molecular formula

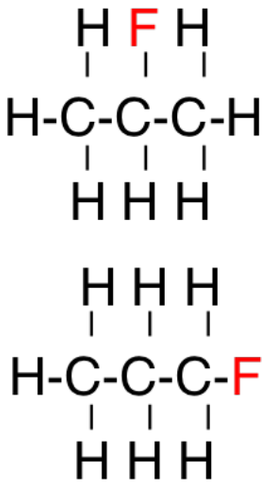


Constitutional (structural) isomers

Stereoisomers (spatial isomers)

Isomers

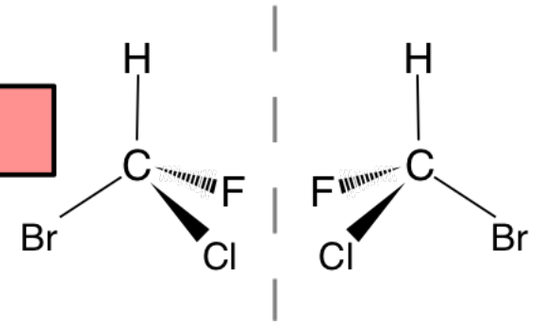
share one molecular formula



Constitutional (structural) isomers

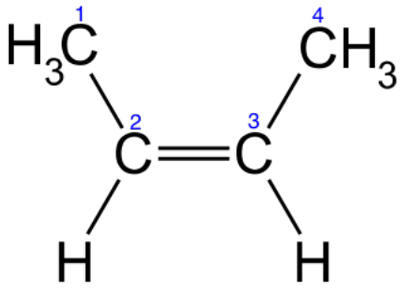
Stereoisomers (spatial isomers)

Diastereomers

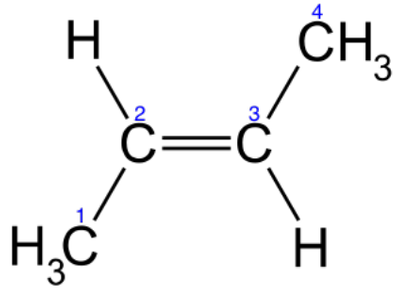


Enantiomers

cis/trans isomers



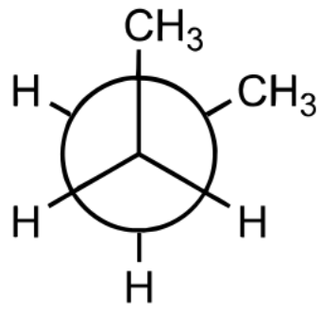
cis-2-butene

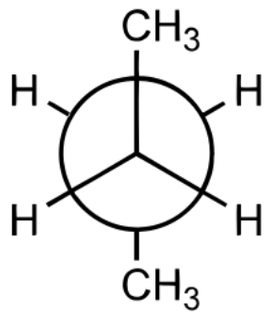


trans-2-butene

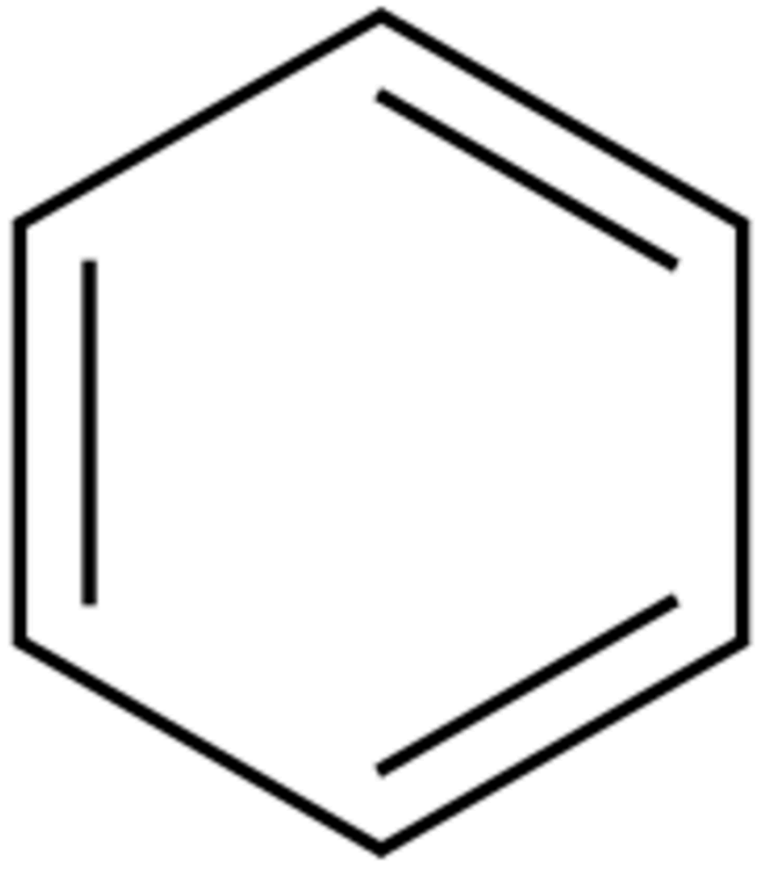
Conformers

Rotamers

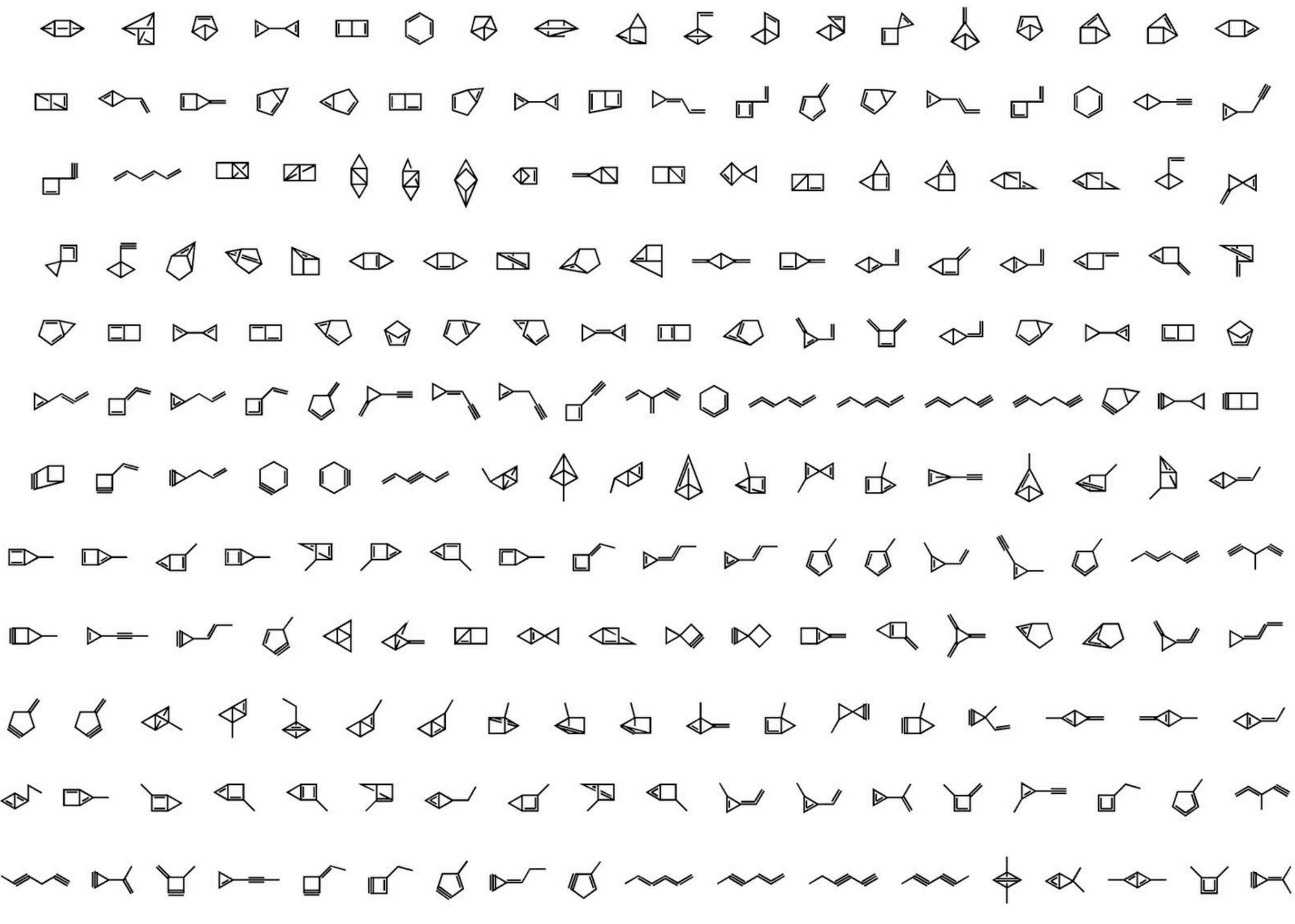




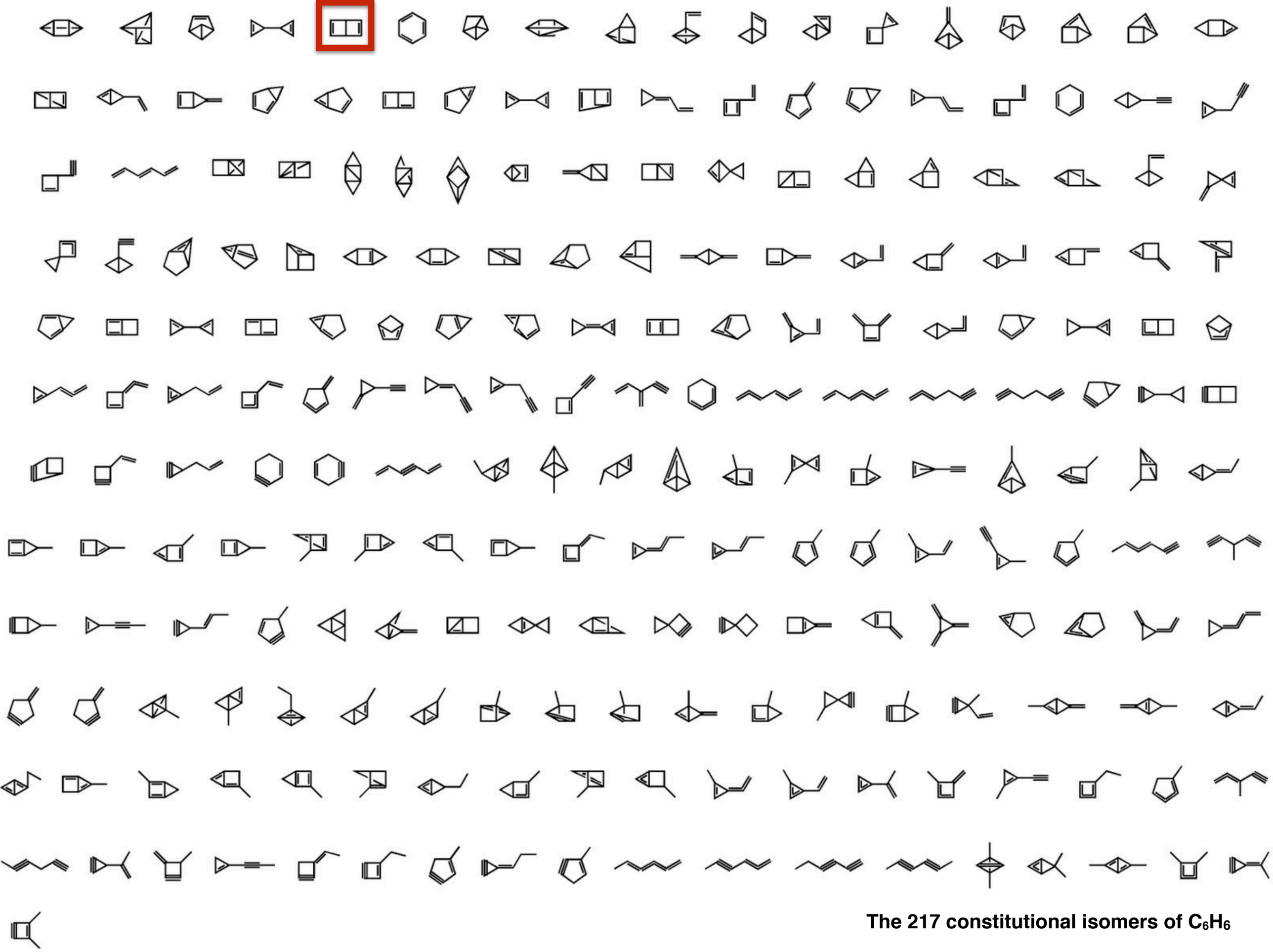
How many constitutional isomers are we looking at?



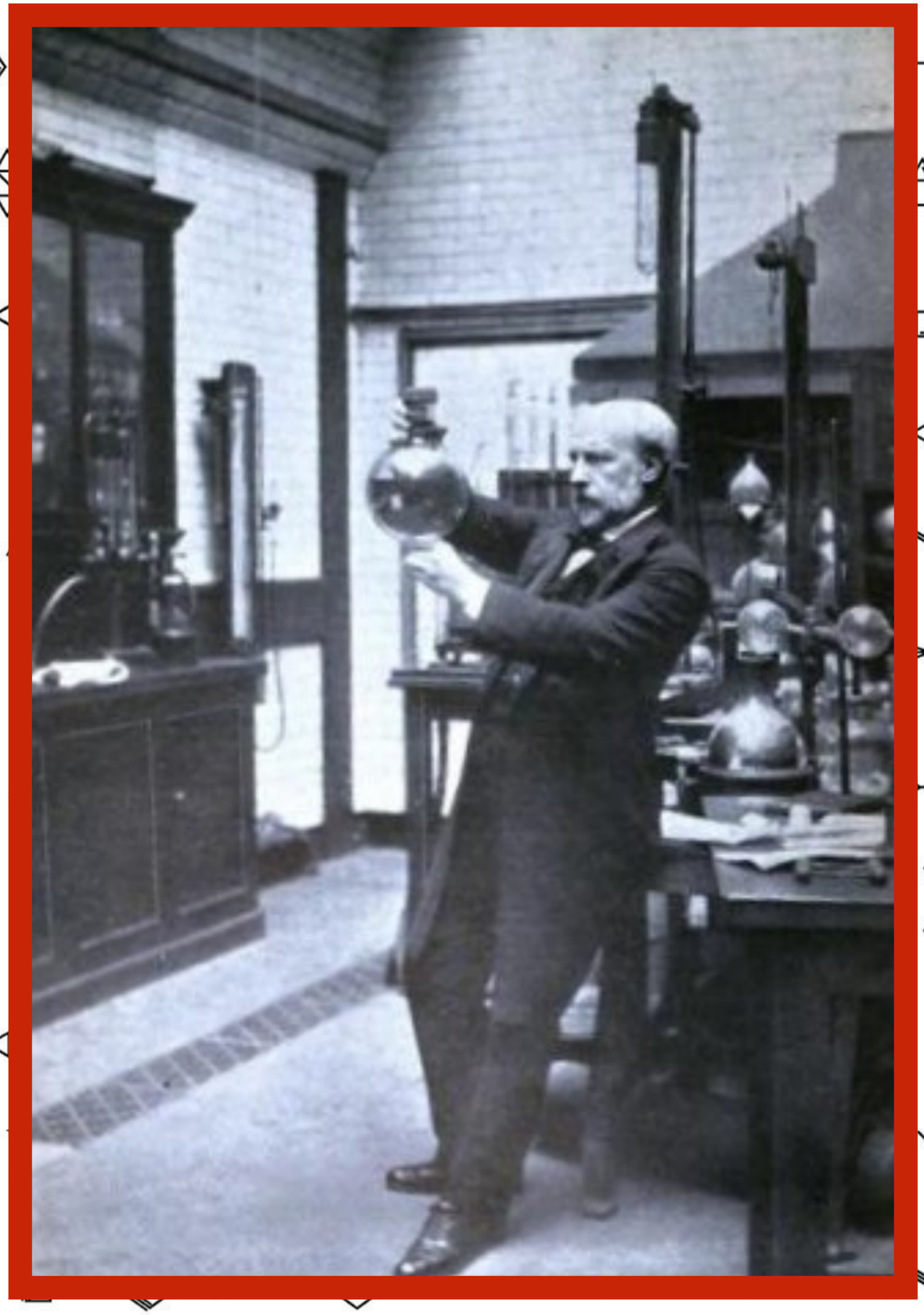
C_6H_6



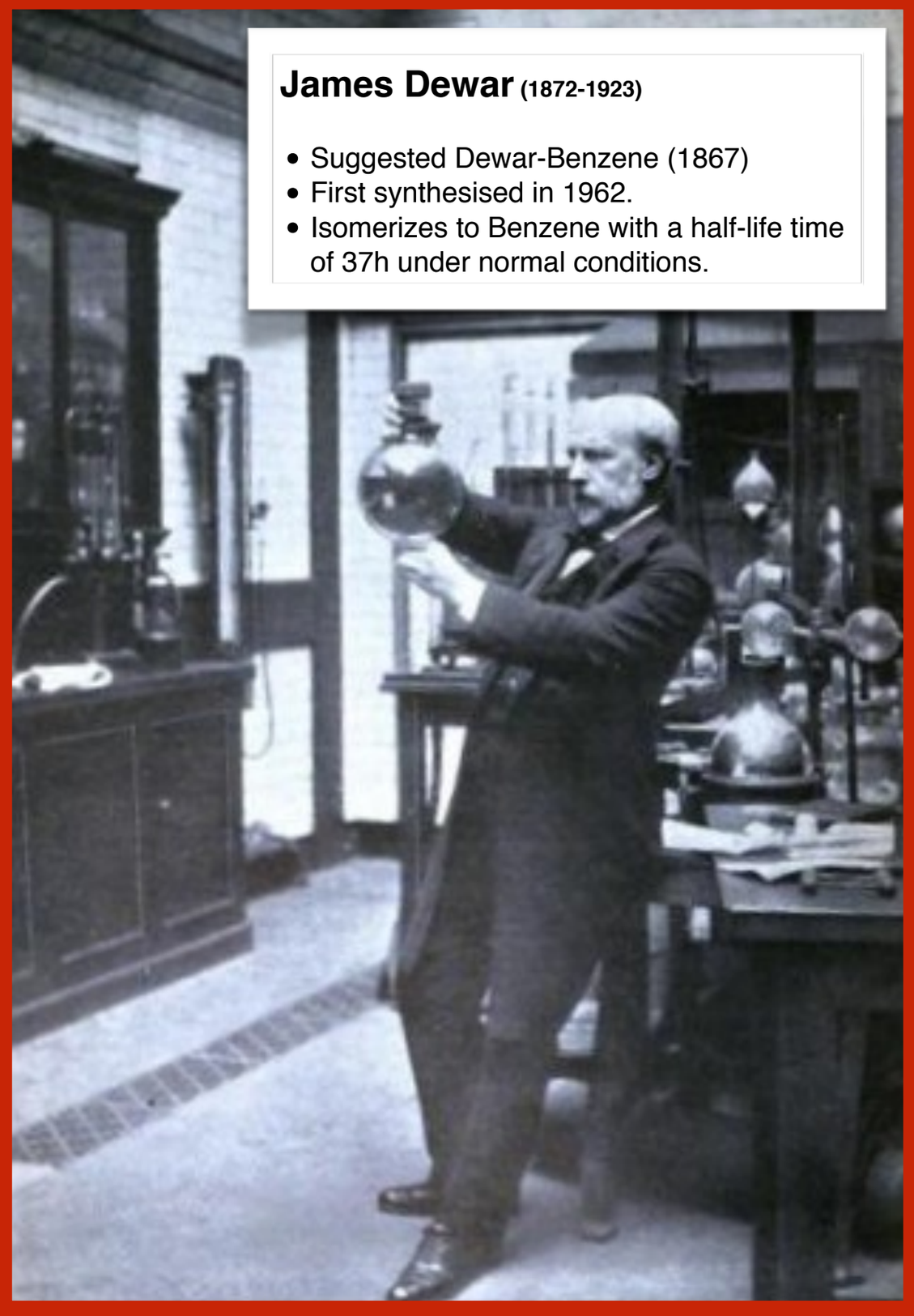
The 217 constitutional isomers of C_6H_6



The 217 constitutional isomers of C_6H_6



The 217 constitutional isomers of C_6H_6



James Dewar (1872-1923)

- Suggested Dewar-Benzene (1867)
- First synthesised in 1962.
- Isomerizes to Benzene with a half-life time of 37h under normal conditions.

The 217 constitutional isomers of C_6H_6

How many constitutional isomers are we looking at?

$C_{10}H_{16}$

$C_{13}H_{16}O_3$

$C_{30}H_{48}O_2$

How many constitutional isomers are we looking at?

$C_{10}H_{16}$ 24938 constitutional isomers

$C_{13}H_{16}O_3$

$C_{30}H_{48}O_2$

How many constitutional isomers are we looking at?

$C_{10}H_{16}$	24938 constitutional isomers
$C_{13}H_{16}O_3$	> 2,000,000,000 constitutional isomers
$C_{30}H_{48}O_2$	

How many constitutional isomers are we looking at?

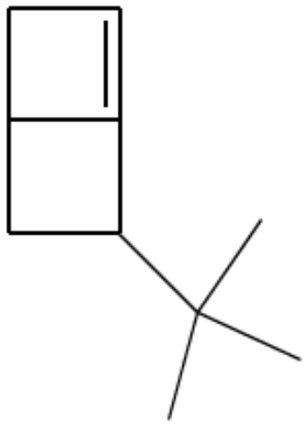
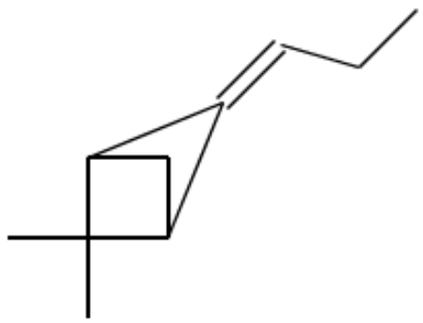
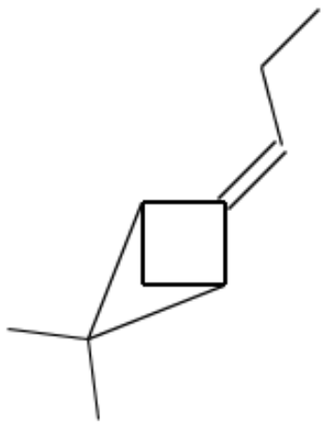
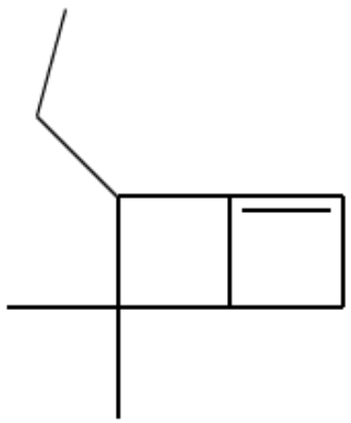
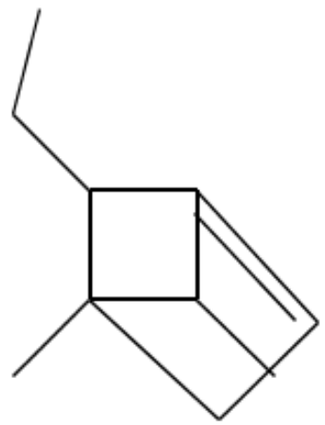
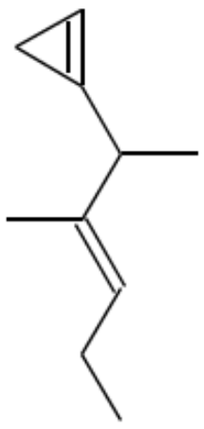
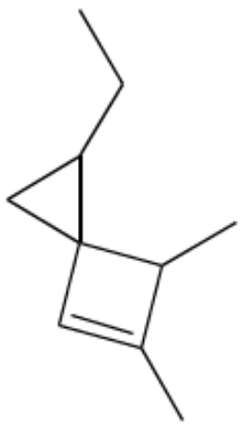
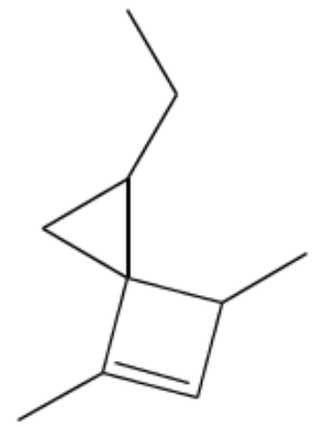
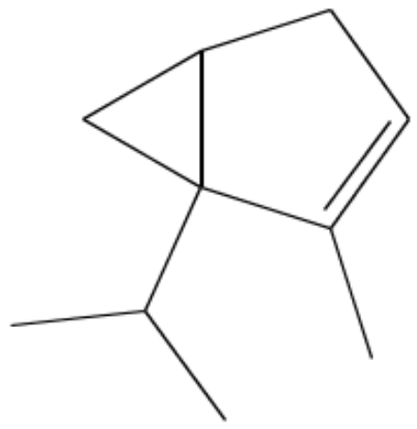
$C_{10}H_{16}$	24938 constitutional isomers
$C_{13}H_{16}O_3$	> 2,000,000,000 constitutional isomers
$C_{30}H_{48}O_2$	>> 10^{12} constitutional isomers

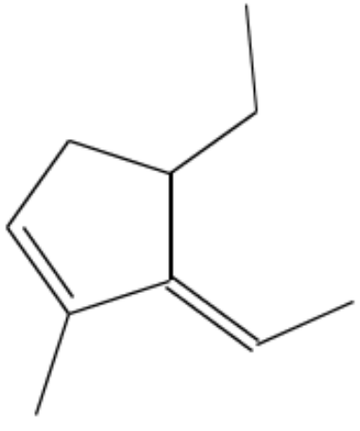
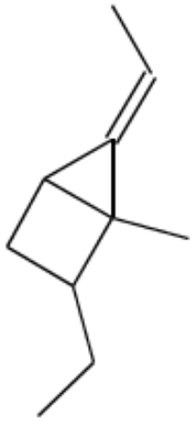
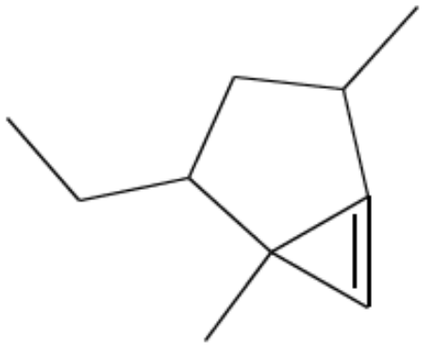
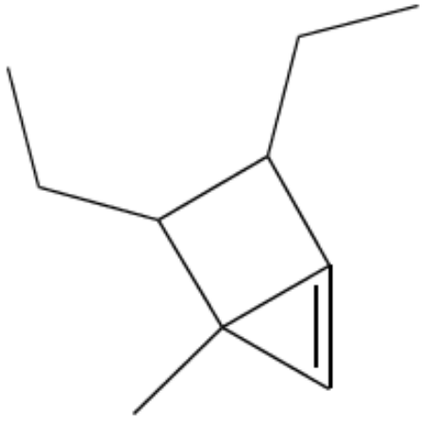
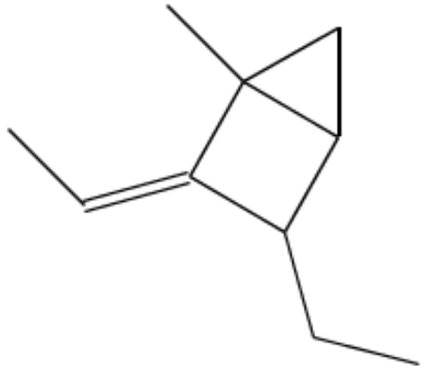
Assume you had a computational tool to make all isomers.

Structure Generators

are **computer programs** to generate **all constitutional isomers** which adhere to a given set of input constraints, such as a molecular formula

Constitutional Isomers of $C_{10}H_{16}$

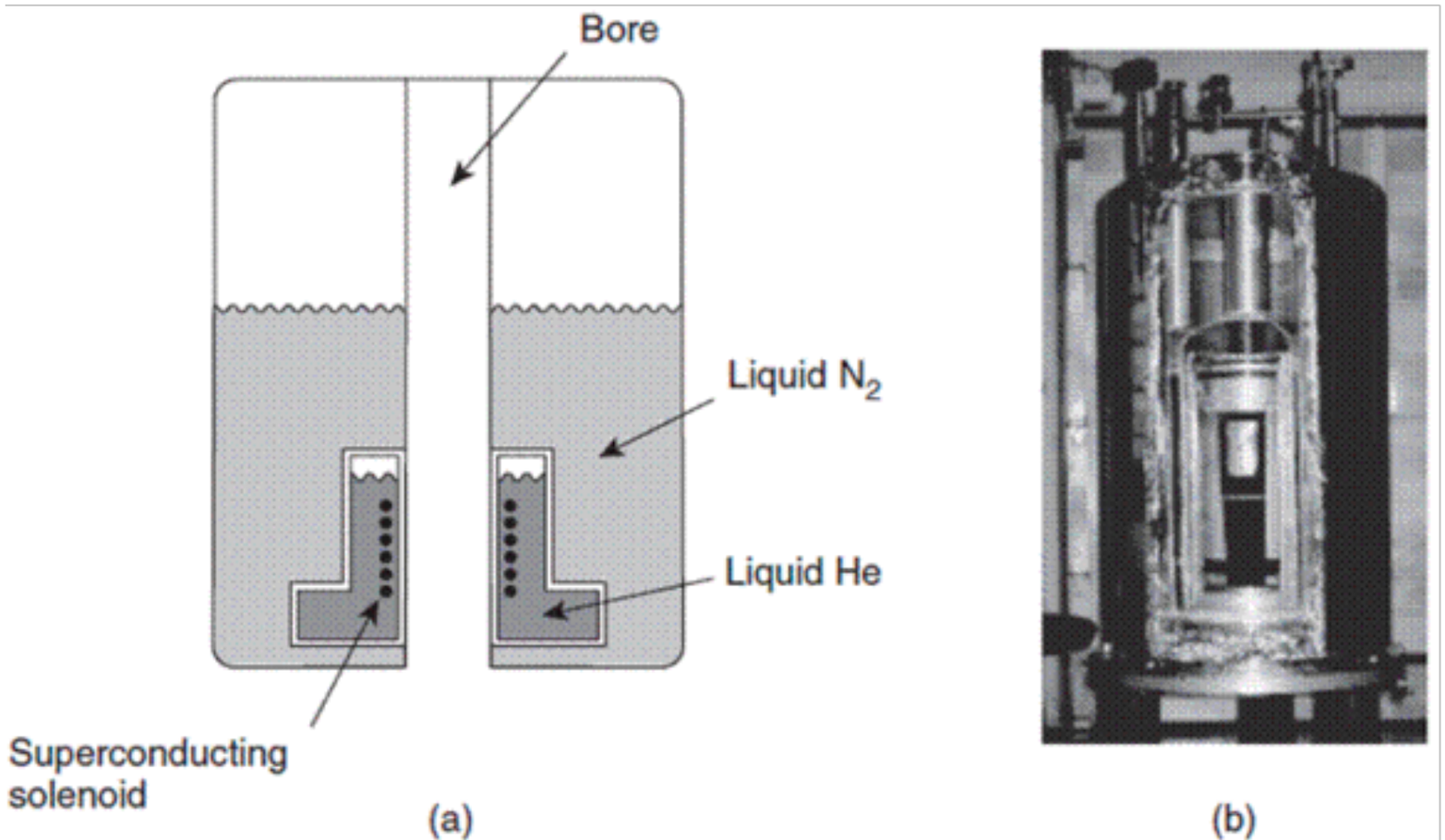






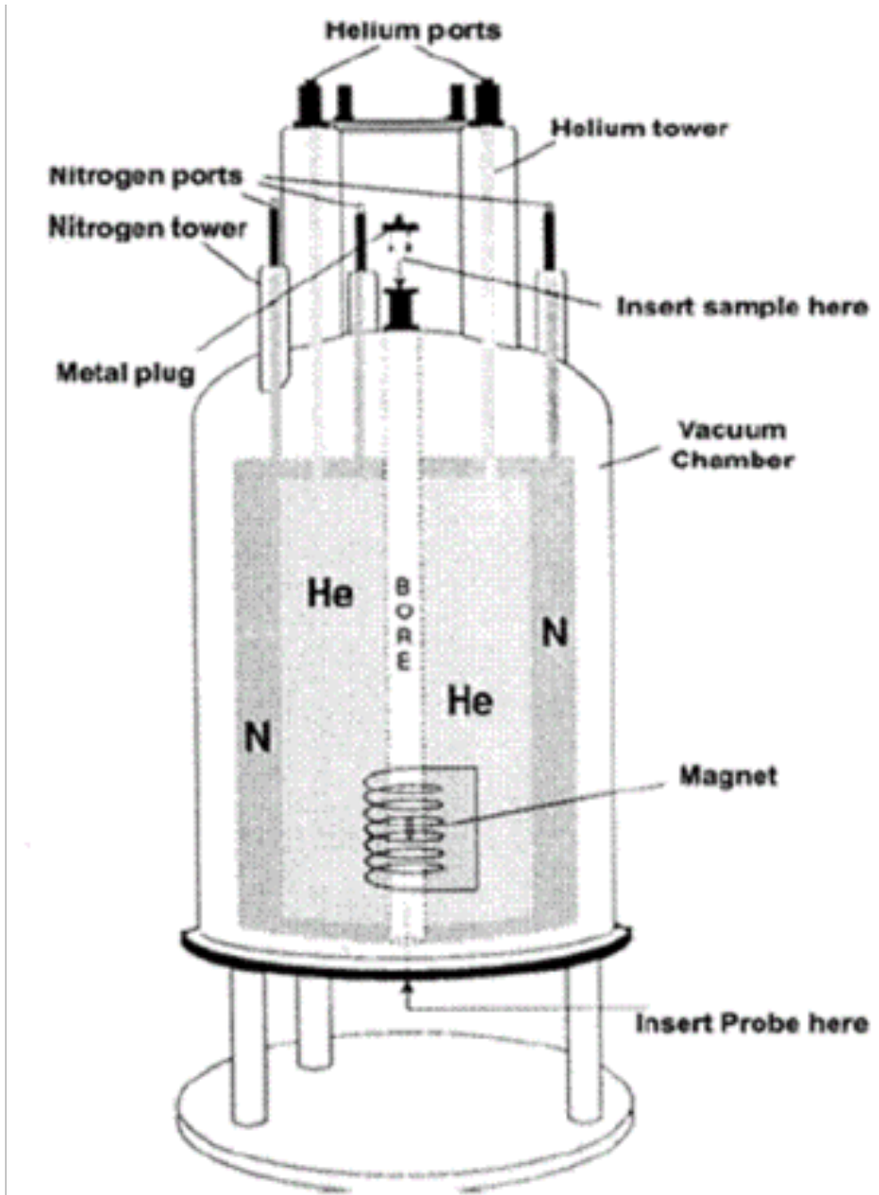
Computer-Assisted Structure Elucidation with **2D** NMR

The only viable way for slightly more complex problems upwards.


BRUKER
Ascend™ 800
AVANCE III HD
BRUKER
PLATFORM

Cryomagnet Design

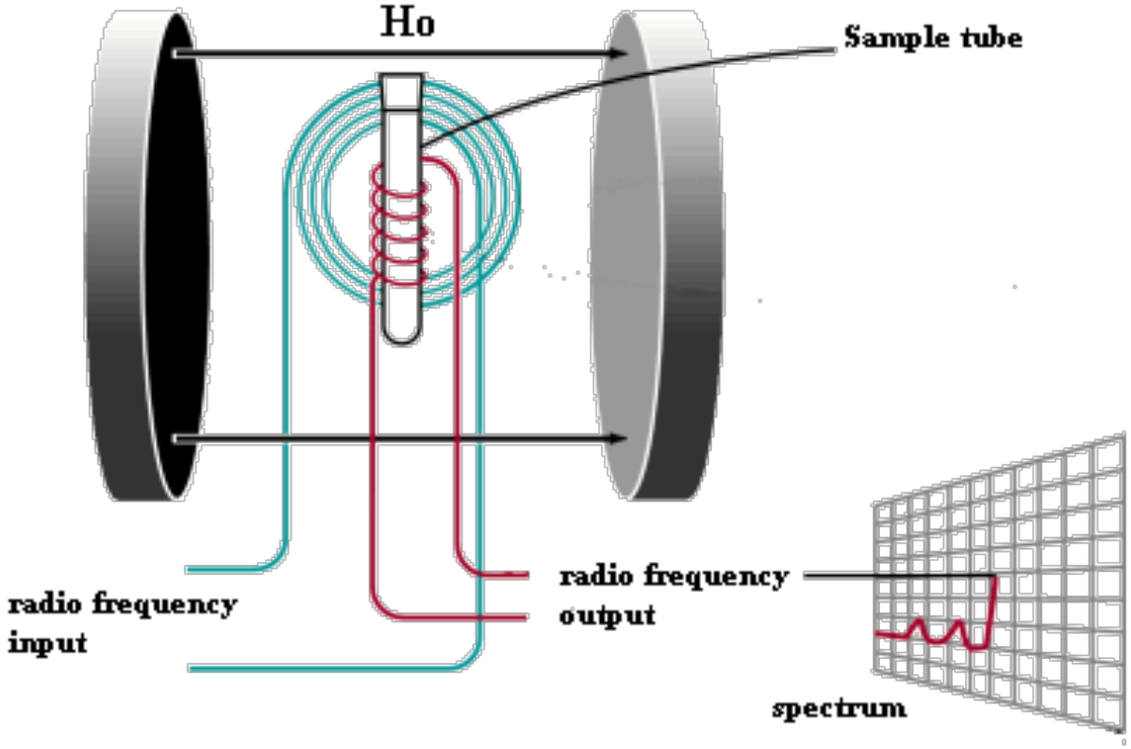




NMR Principle

Odd number of protons and/or neutrons ->
intrinsic magnetic moment and angular momentum (nonzero spin)

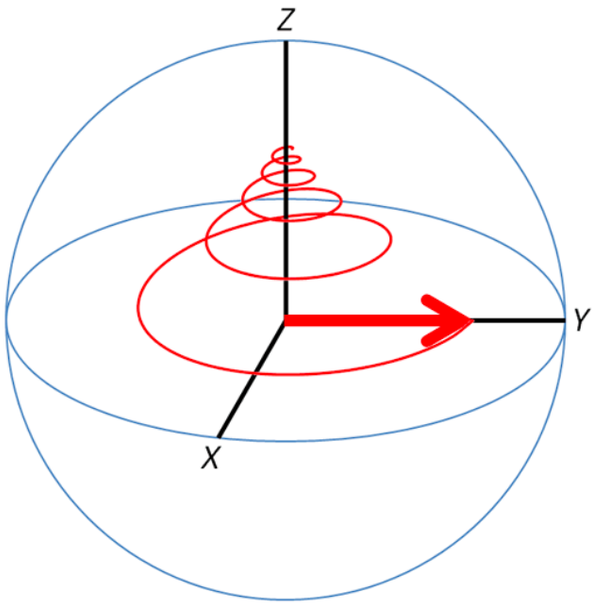
Even numbers of both -> spin of zero.

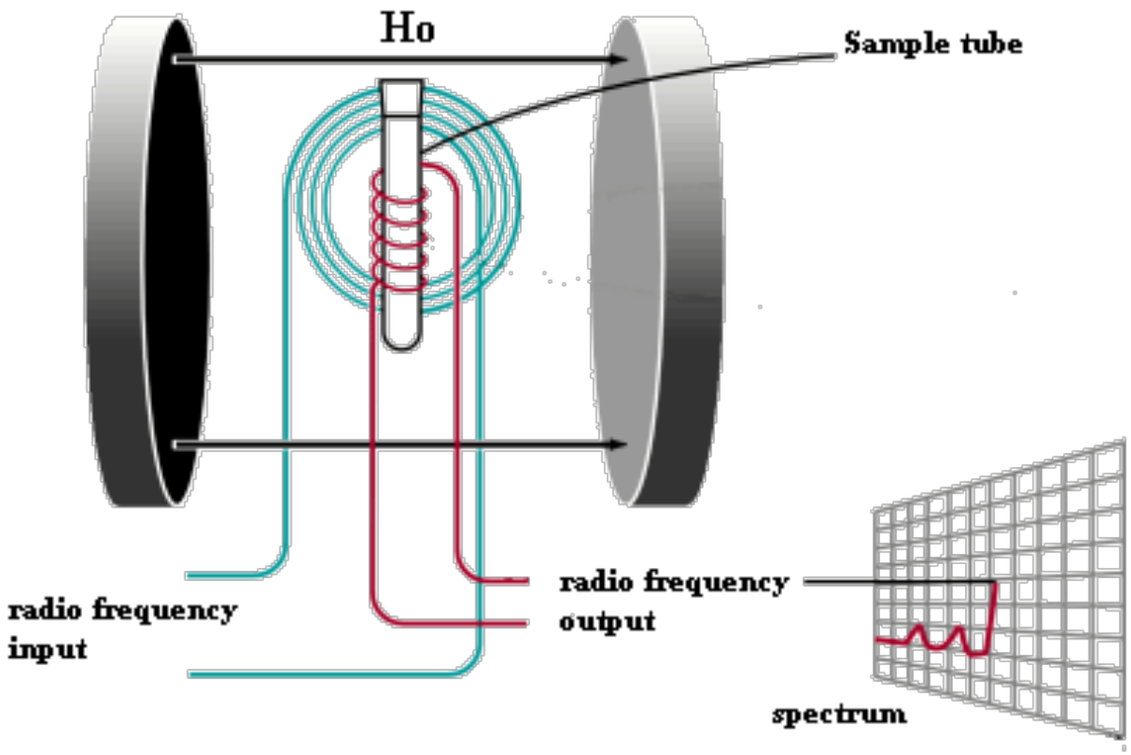


NMR Principle

Odd number of protons and/or neutrons ->
intrinsic magnetic moment and angular momentum (nonzero spin)

Even numbers of both -> spin of zero.

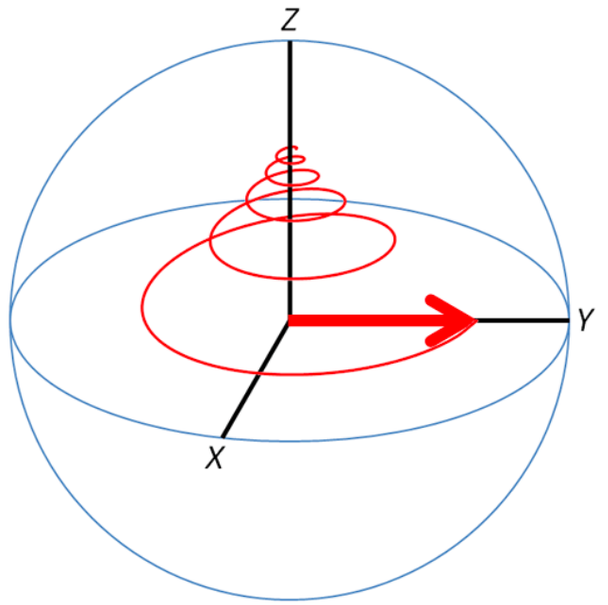


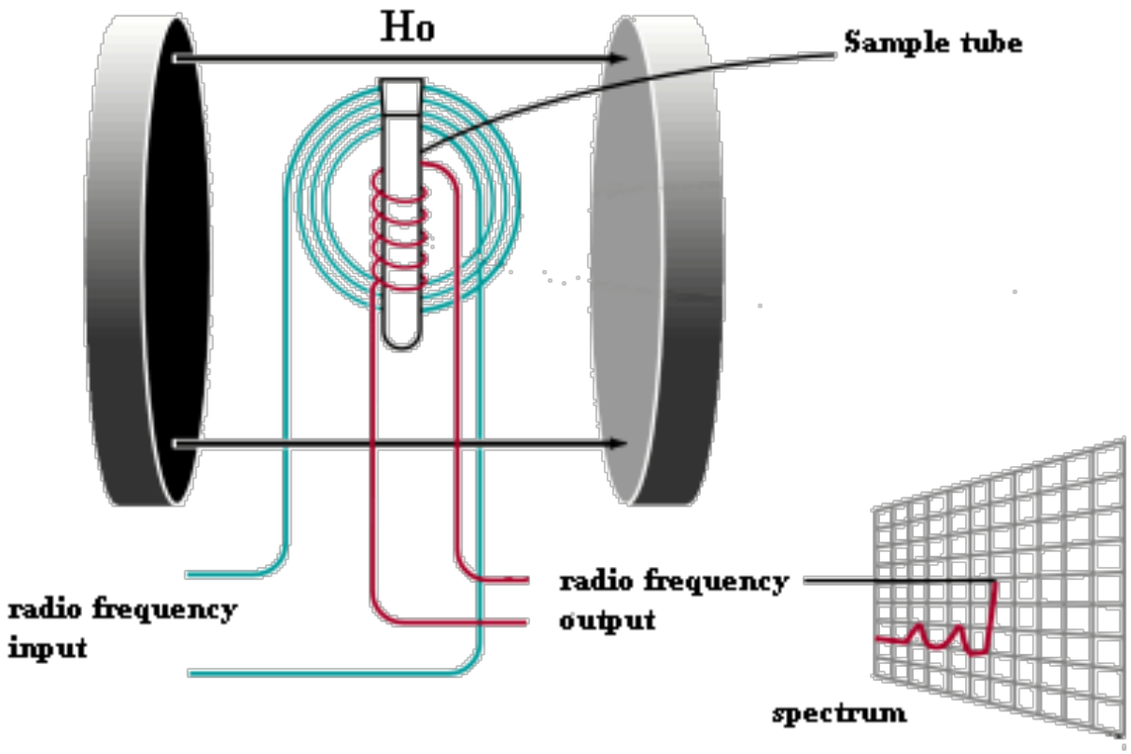


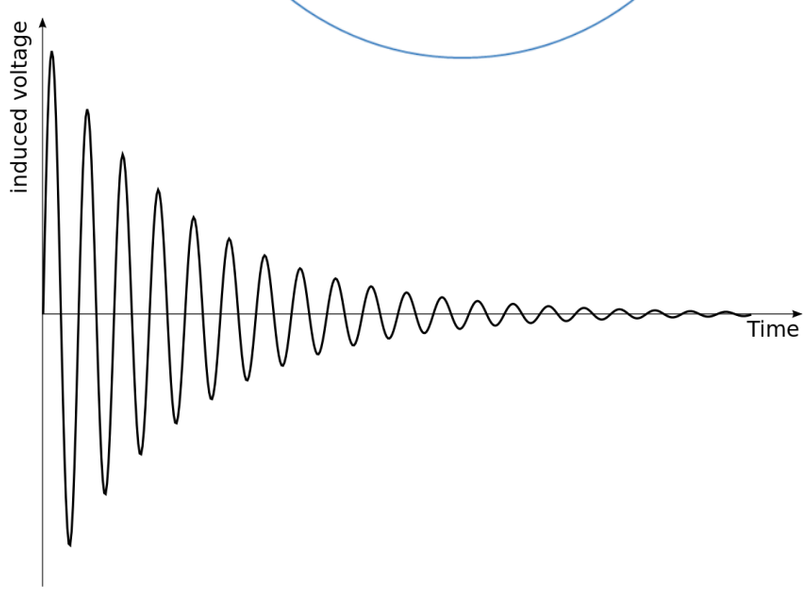
NMR Principle

Odd number of protons and/or neutrons ->
intrinsic magnetic moment and angular momentum (nonzero spin)

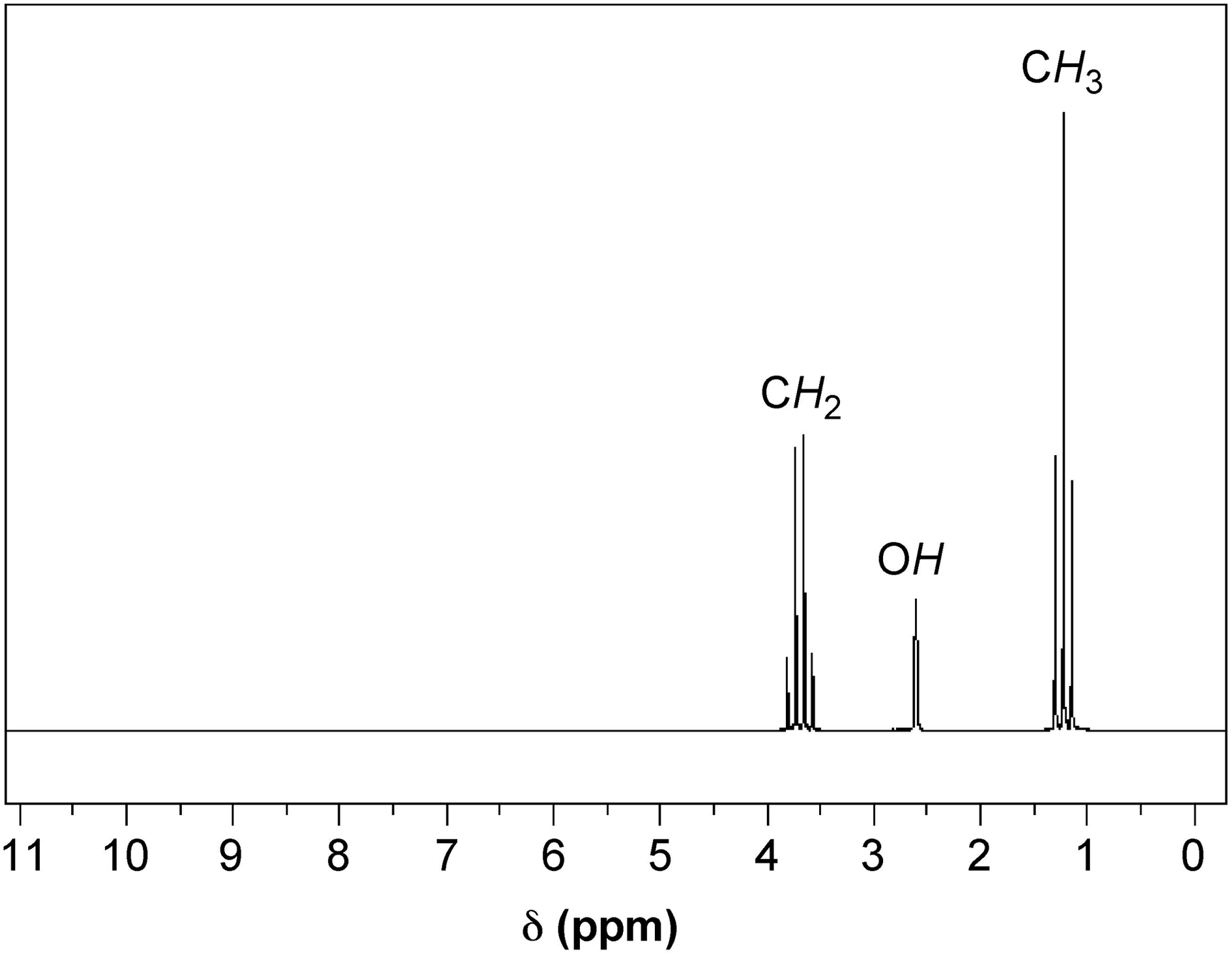
Even numbers of both -> spin of zero.



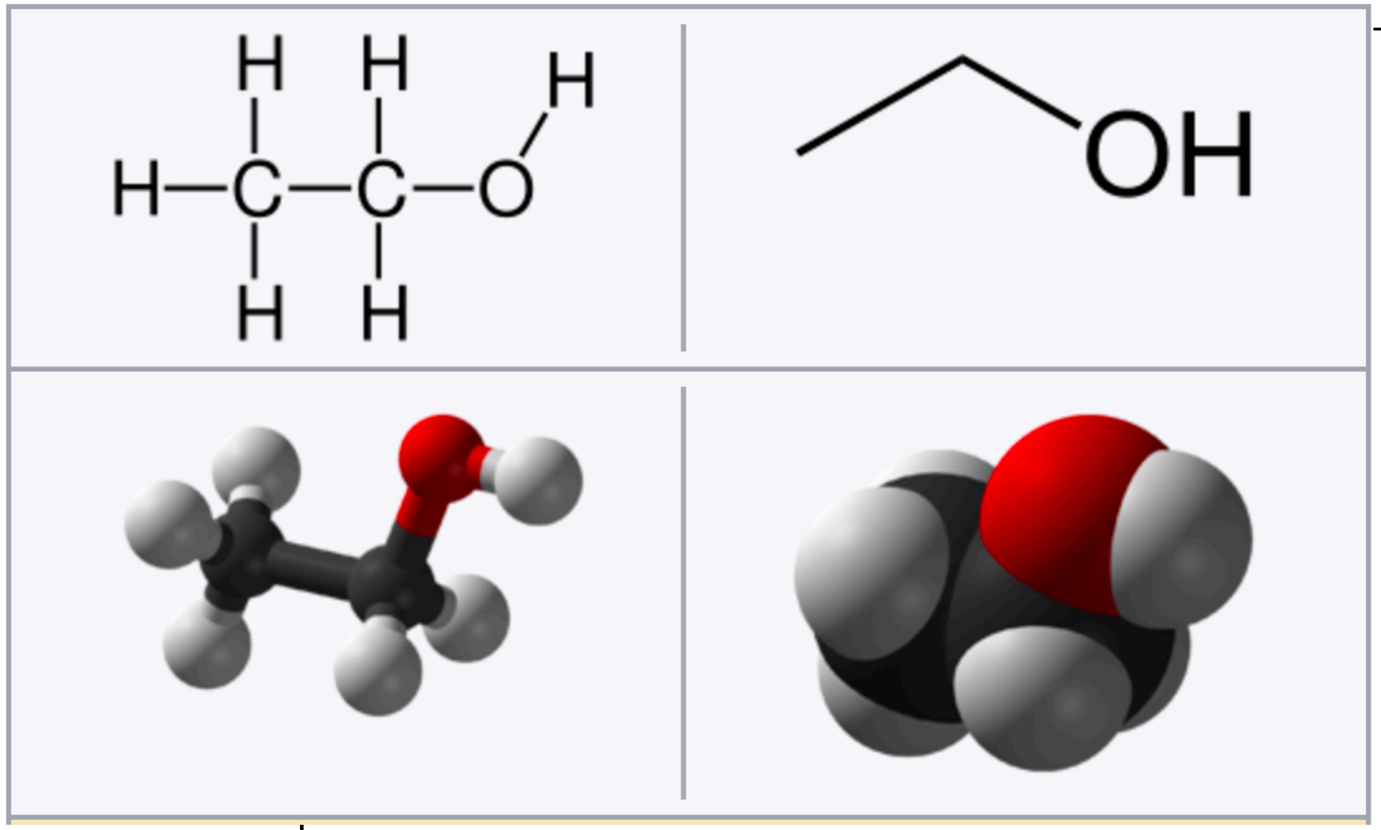




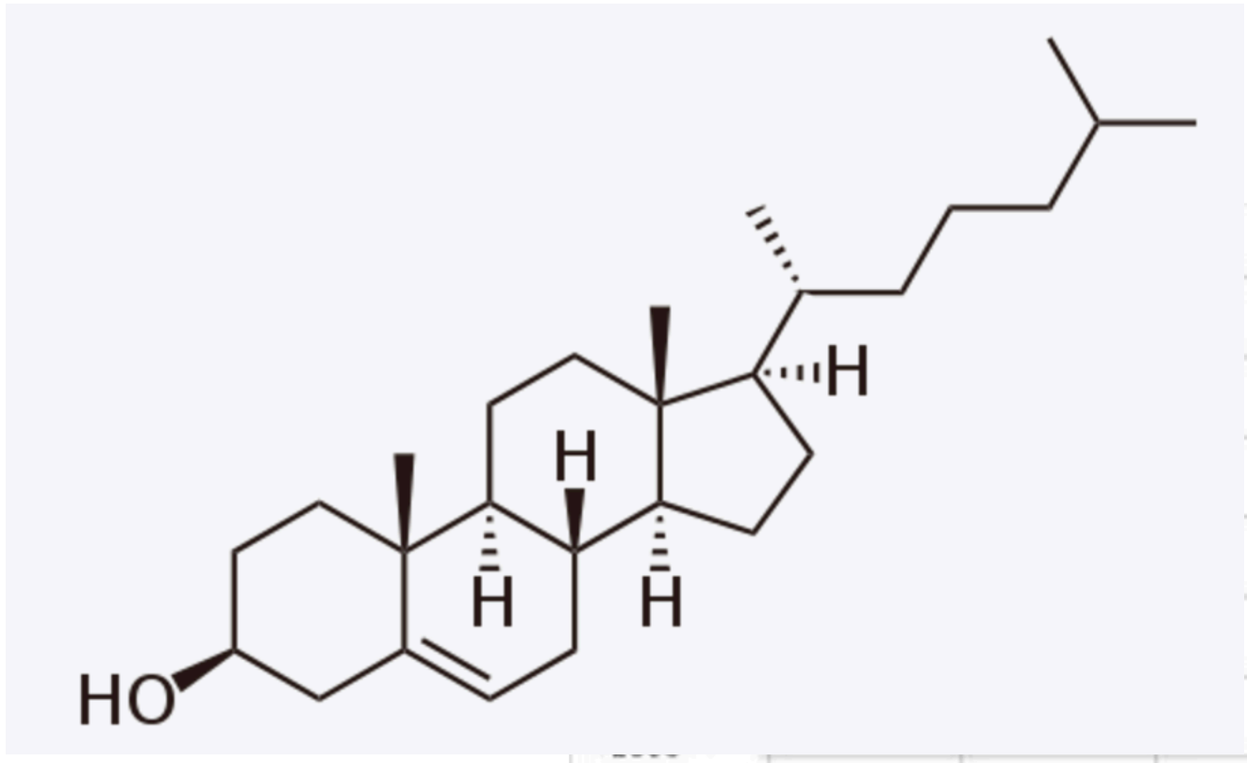
1D Proton NMR



1D Proton NMR

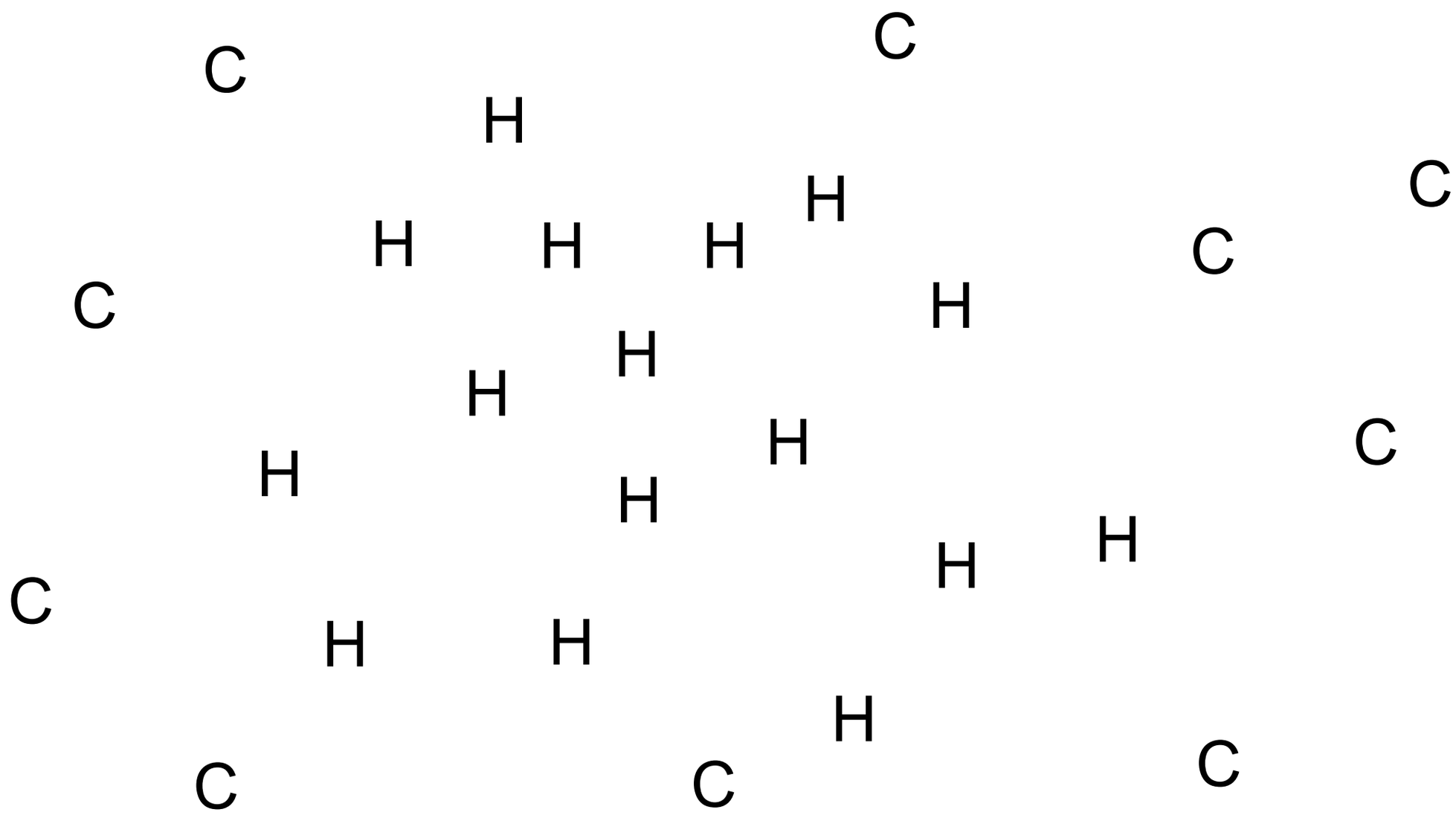


1D Proton NMR



Exploded Pharmacy: HR-MS

yields information about elemental composition, such as $C_{10}H_{16}$

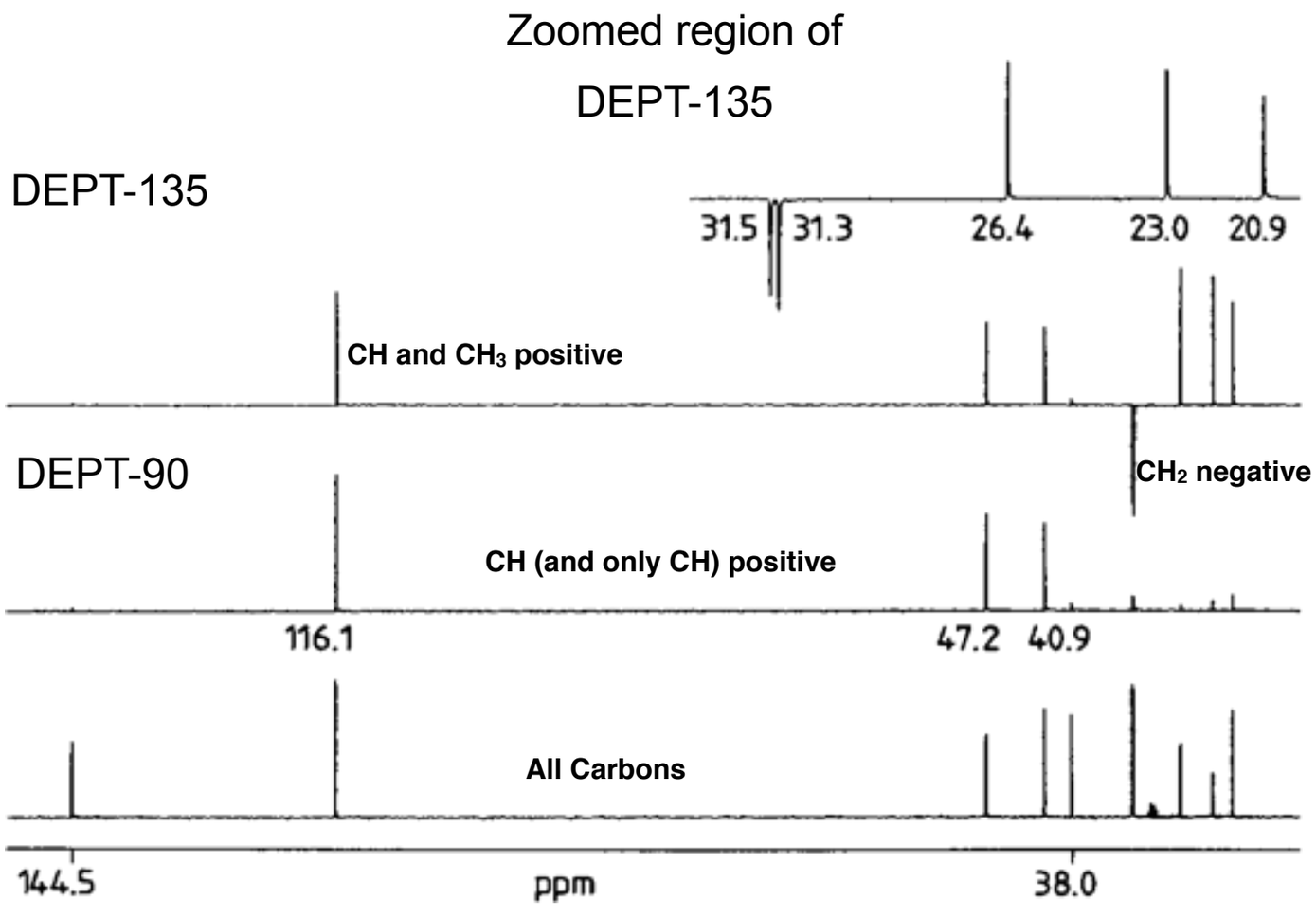


Experiments: J-Couplings: DEPT

Acronyms:

DEPT: Distortionless Enhancement by Polarization Transfer

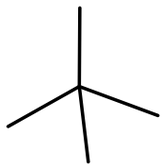
APT: Attached Proton Test



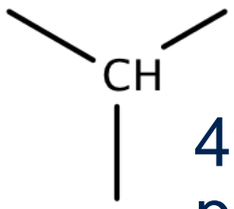
- ^{13}C-detekted 1D-Exp.
- Number of protons attached to each carbon is coded as signal phase
- Combining information from DEPT-135, DEPT-90 and bb-decoupled carbon nmr yields a complete list of carbon fragments in the molecule.

Time required: 1-5 min

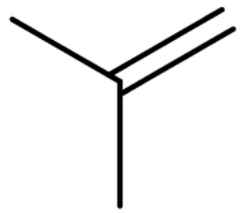
After evaluation of DEPT experiment (or multiplicity edited HSQC) heavy atoms are labeled with chemical shifts and number of attached hydrogen atoms.



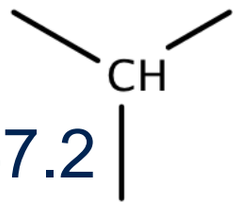
38.0 ppm



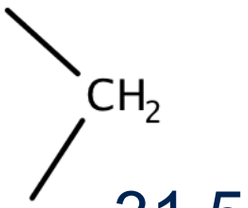
40.9 ppm



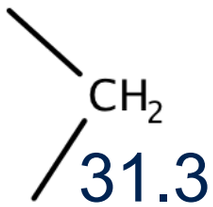
144.5 ppm



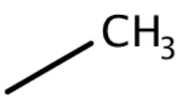
47.2 ppm



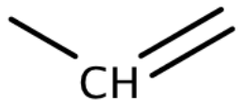
31.5 ppm



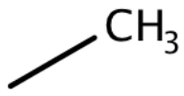
31.3 ppm



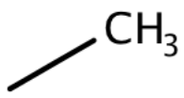
20.9 ppm



116.1 ppm



23.0 ppm



26.4 ppm

Experiments: J-Couplings: HSQC

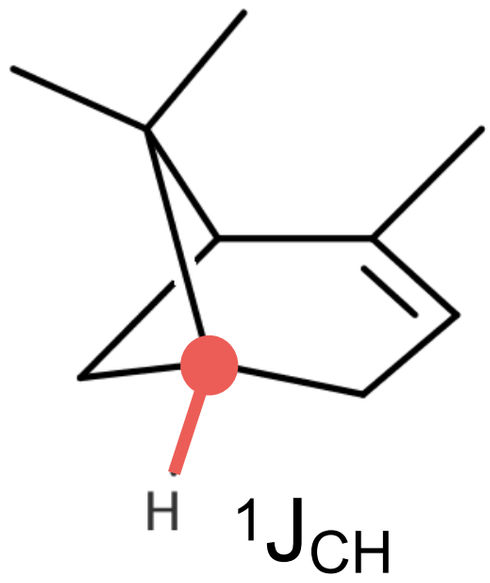
alias HMQC, CH-COSY, HETCOR

Acronyms:

HSQC: Heteronuclear Single Quantum Coherence

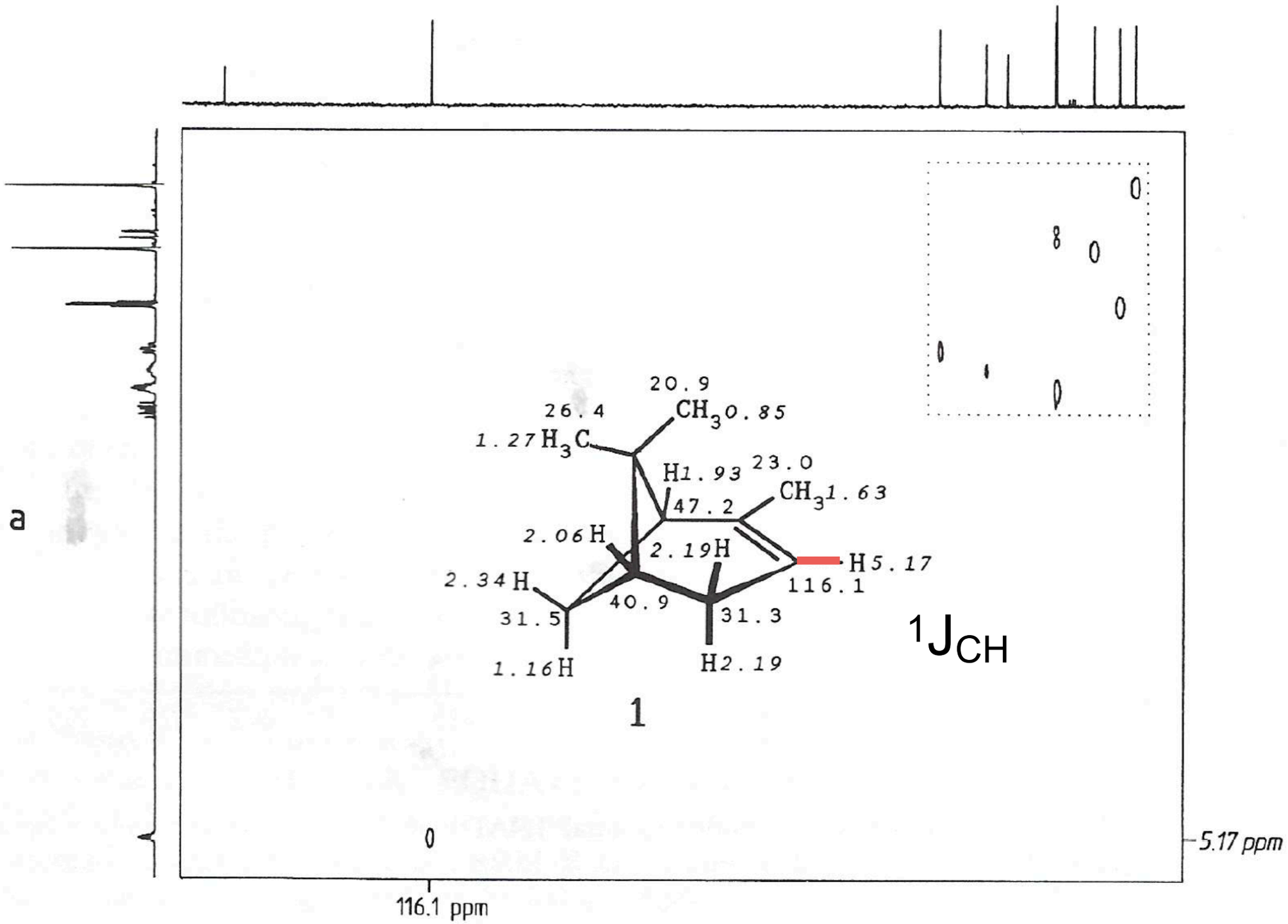
HMQC: Heteronuclear Multiple Quantum Coherence

- Cross-Signals in HSQC-diagrams caused by CH-Couplings via **one** CH-bond($^1J_{CH}$, ~140 Hz).
- Experiment yields list of pairs of directly bonded carbon and hydrogen atoms.



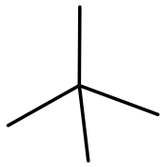
Experiments: J-Couplings: HSQC

alias HMQC, CH-COSY, HETCOR



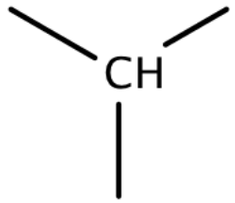
After evaluation of HSQC, our carbon atoms are also labled with the proton shifts of their directly attached protons

38.0 ppm

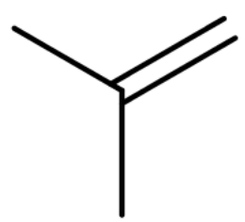


2.06 ppm

40.9 ppm

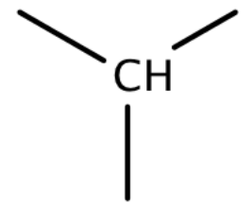


144.5 ppm

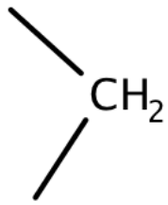


1.93 ppm

47.2 ppm

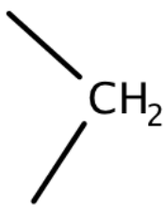


31.5 ppm



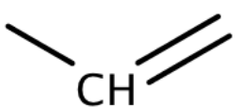
1.16 ppm

2.34 ppm



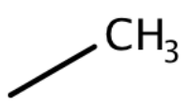
31.3 ppm

2.19 ppm



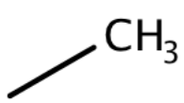
116.1 ppm

5.17 ppm



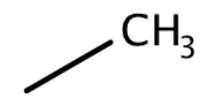
23.0 ppm

1.63 ppm



26.4 ppm

1.27 ppm



20.9 ppm

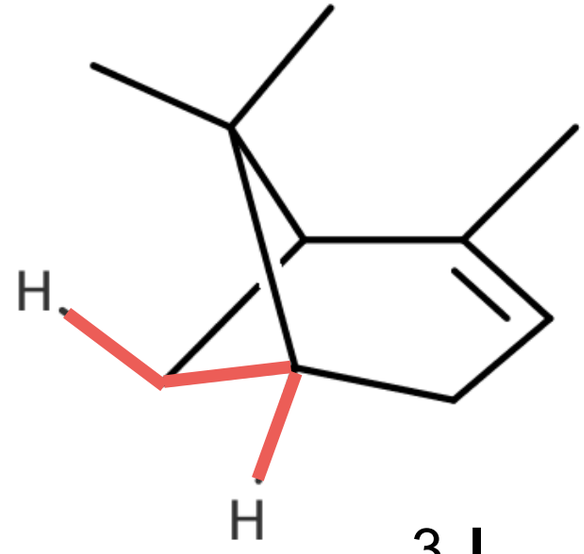
0.85 ppm

Experiments: J-Couplings: HH-COSY

Acronyms:

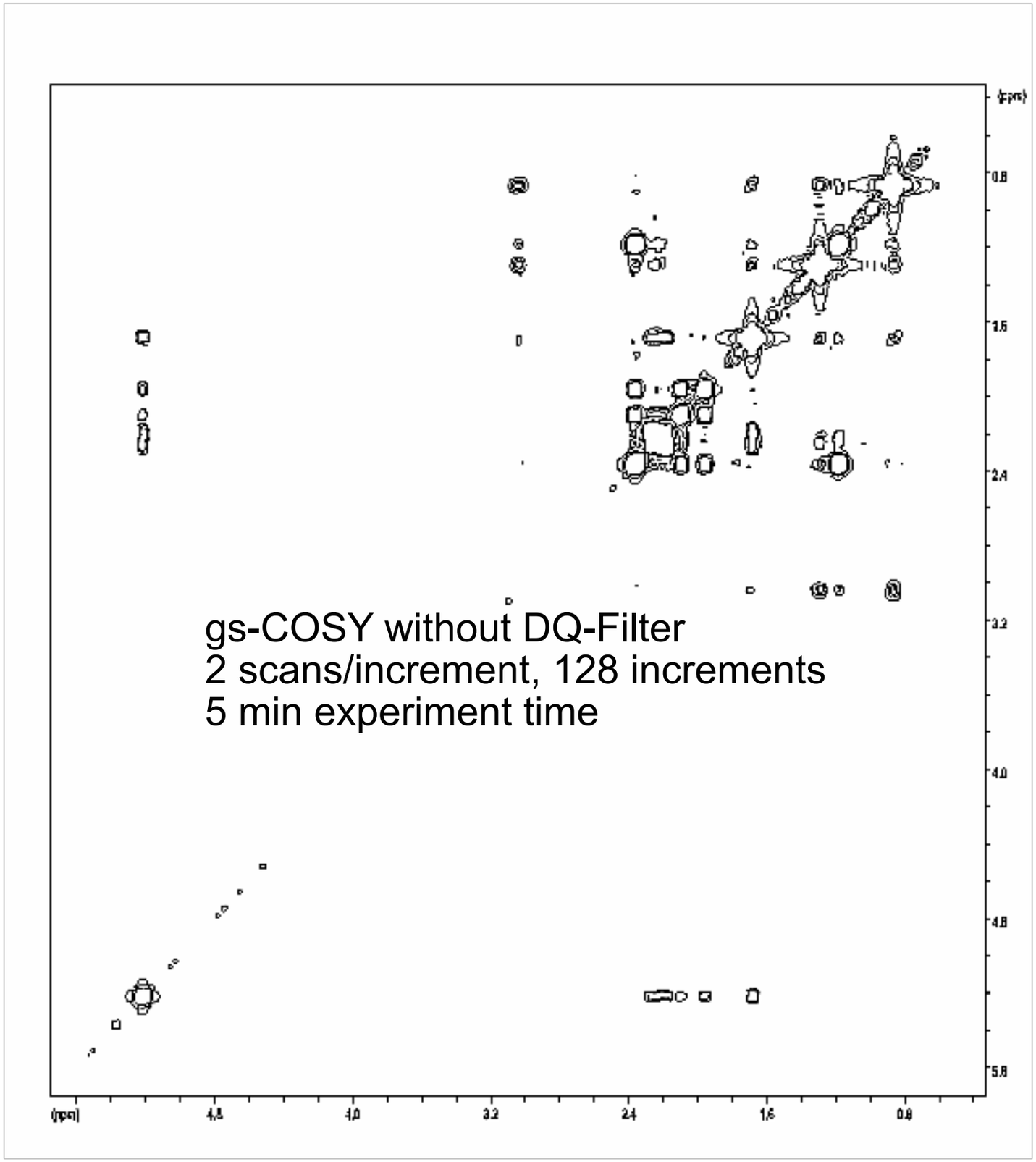
DQF-COSY: Double Quantum Filtered COrrelation Spectroscopy

- $^{3}J_{HH}$-Couplings (and a few others, unfortunately)
- Proton-rich skeletons might be elucidated just by HHCOSY und HSQC
- Problem: Quarternary carbons, heteroatoms in the skeleton

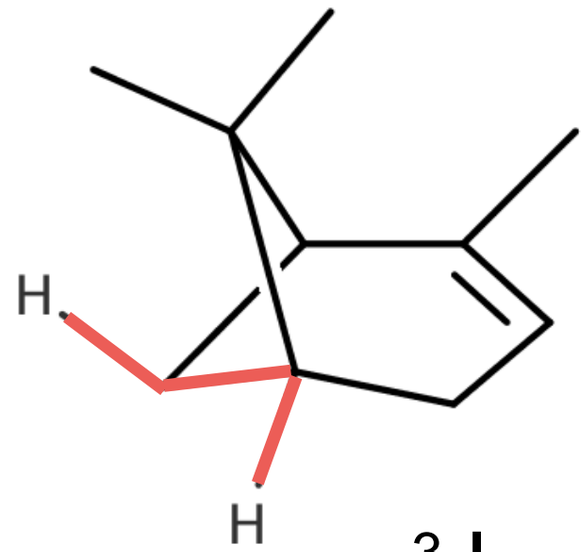


$^{3}J_{HH}$

Experiments: J-Couplings: HH-COSY

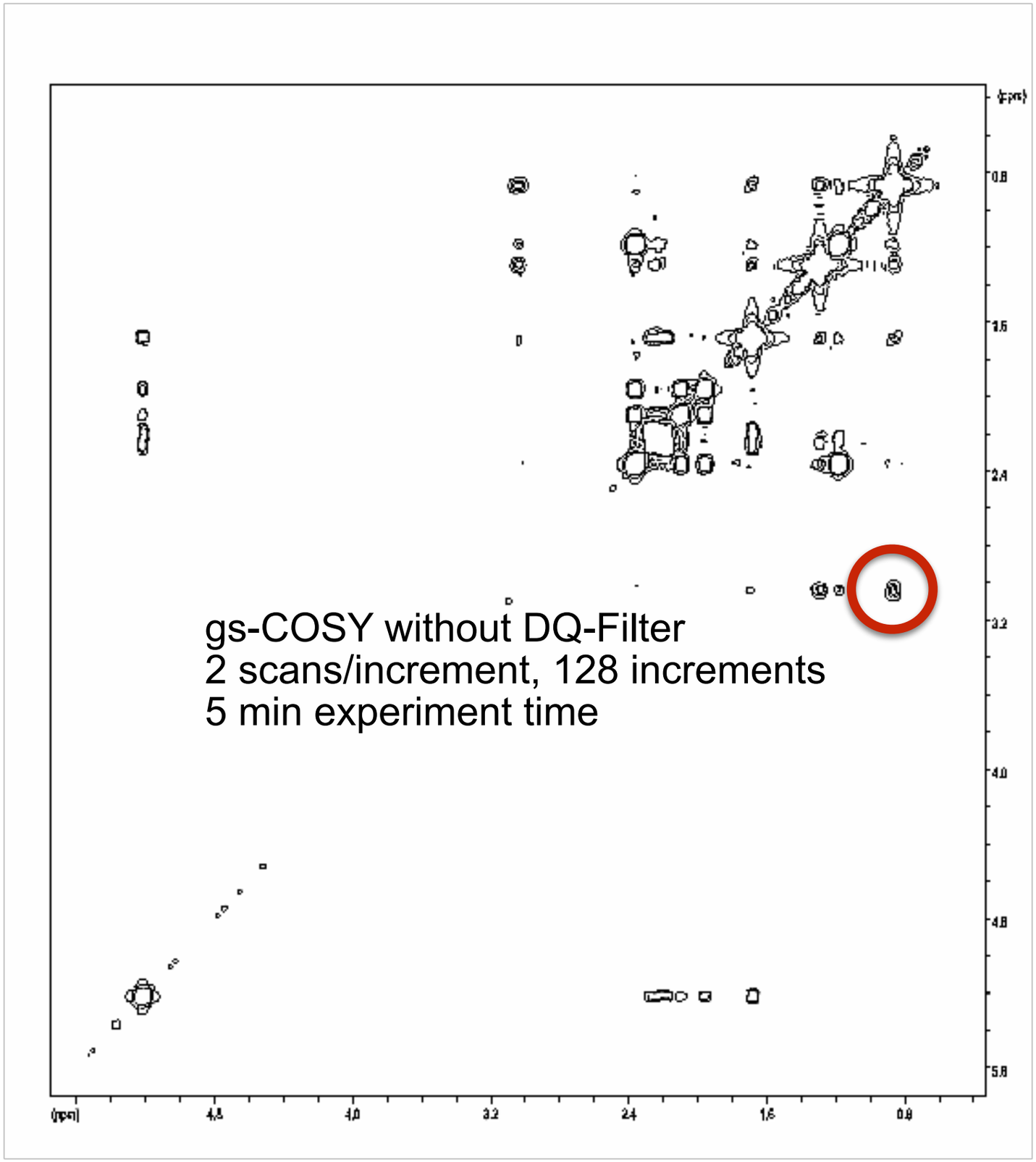


gs-COSY without DQ-Filter
2 scans/increment, 128 increments
5 min experiment time

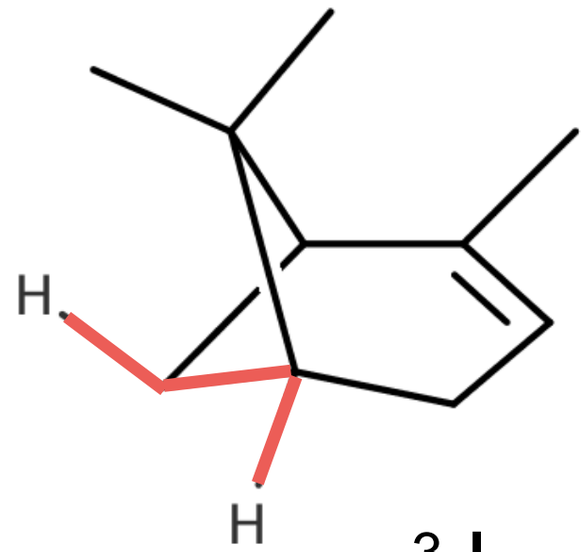


$^3J_{HH}$

Experiments: J-Couplings: HH-COSY



gs-COSY without DQ-Filter
2 scans/increment, 128 increments
5 min experiment time



$^3J_{HH}$

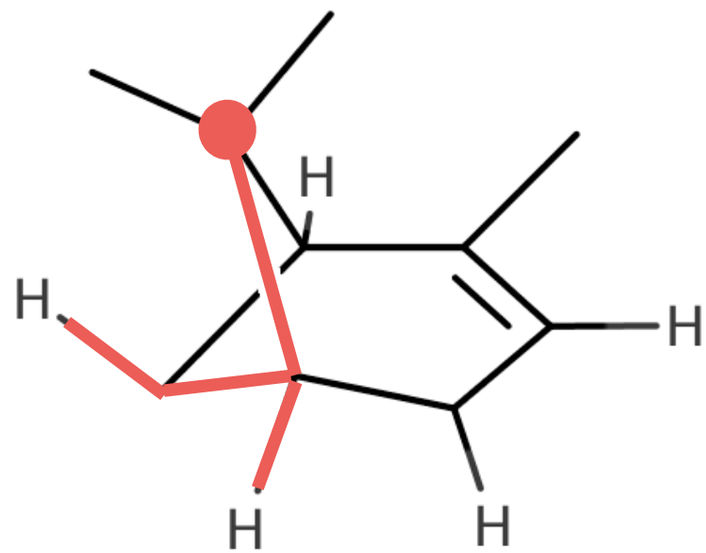
Experiments: J-Couplings: HMBC

Acronyms:

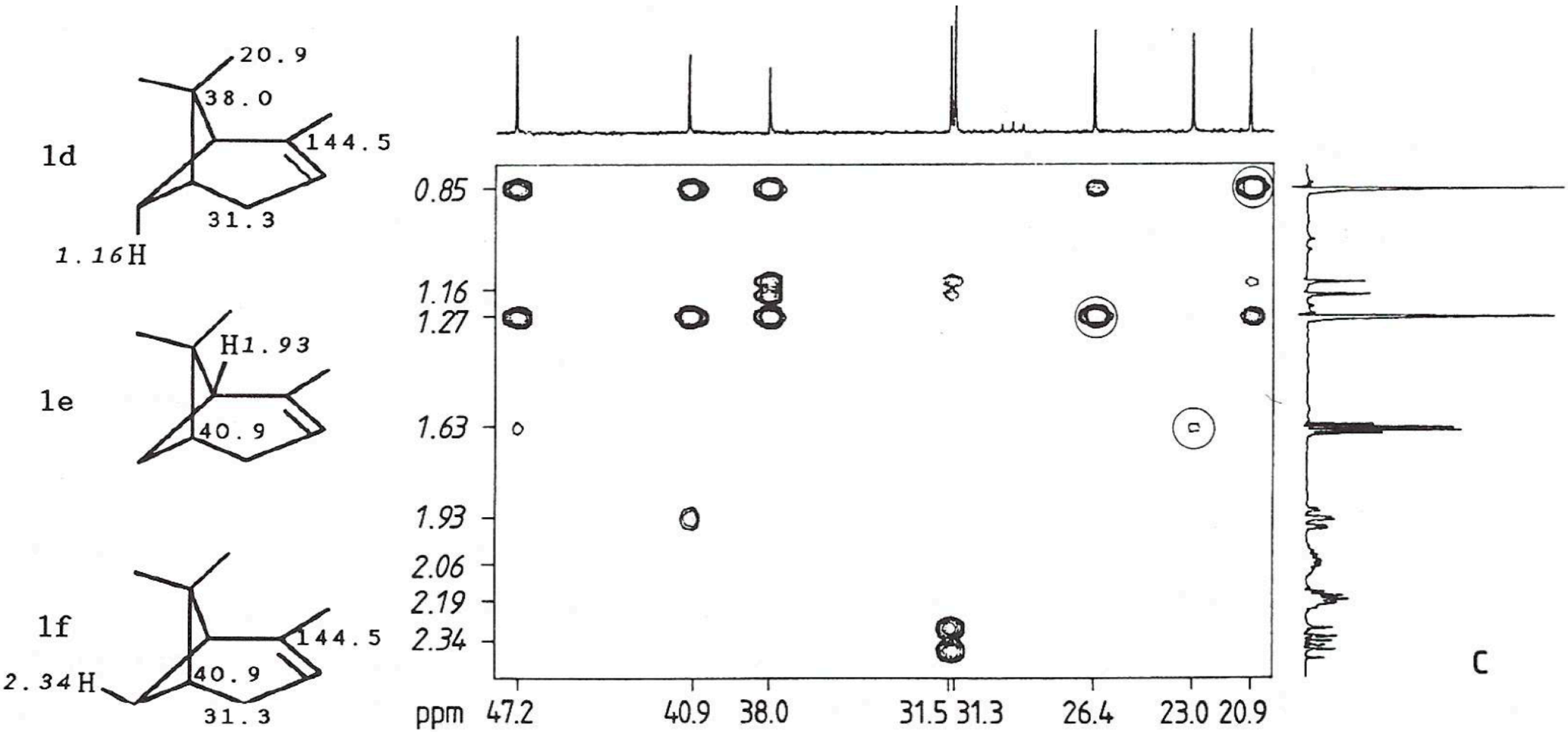
HMBC: Heteronuclear Multiple Bond Coherence

COLOC: COrrelation via LOng range Kopplungen

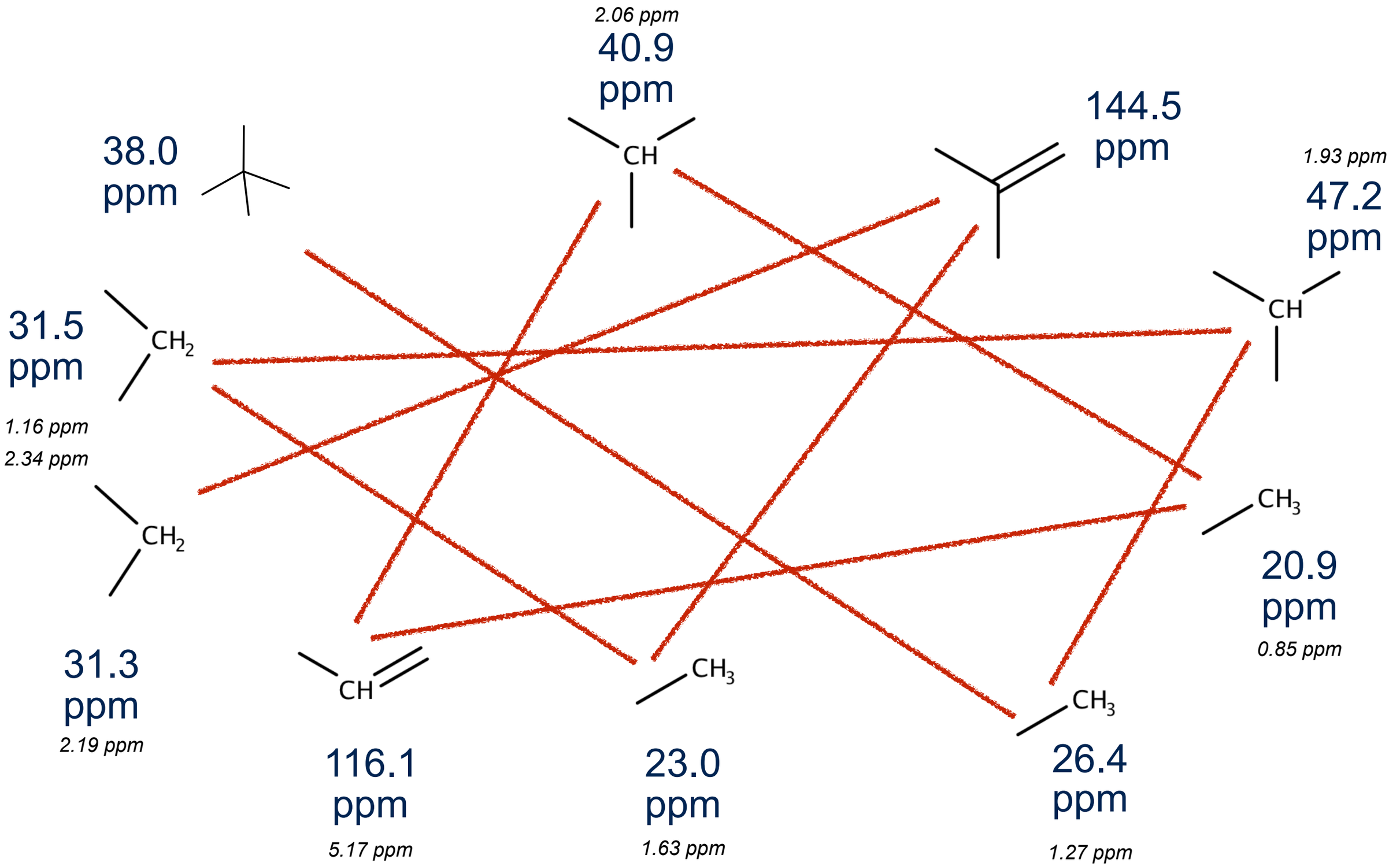
- Cross-signals through scalar couplings between carbon and hydrogen via 2 or 3 bonds ($^2J_{CH}/^3J_{CH}$, ~8 Hz).
- Problem: $^2J_{CH}/^3J_{CH}$ -couplings cannot be distinguished.
- Problem: $^4J_{CH}/^5J_{CH}$ -couplings, which cannot be easily distinguished from $^2J_{CH}/^3J_{CH}$ -couplings



Experiments: J-Couplings-> HMBC

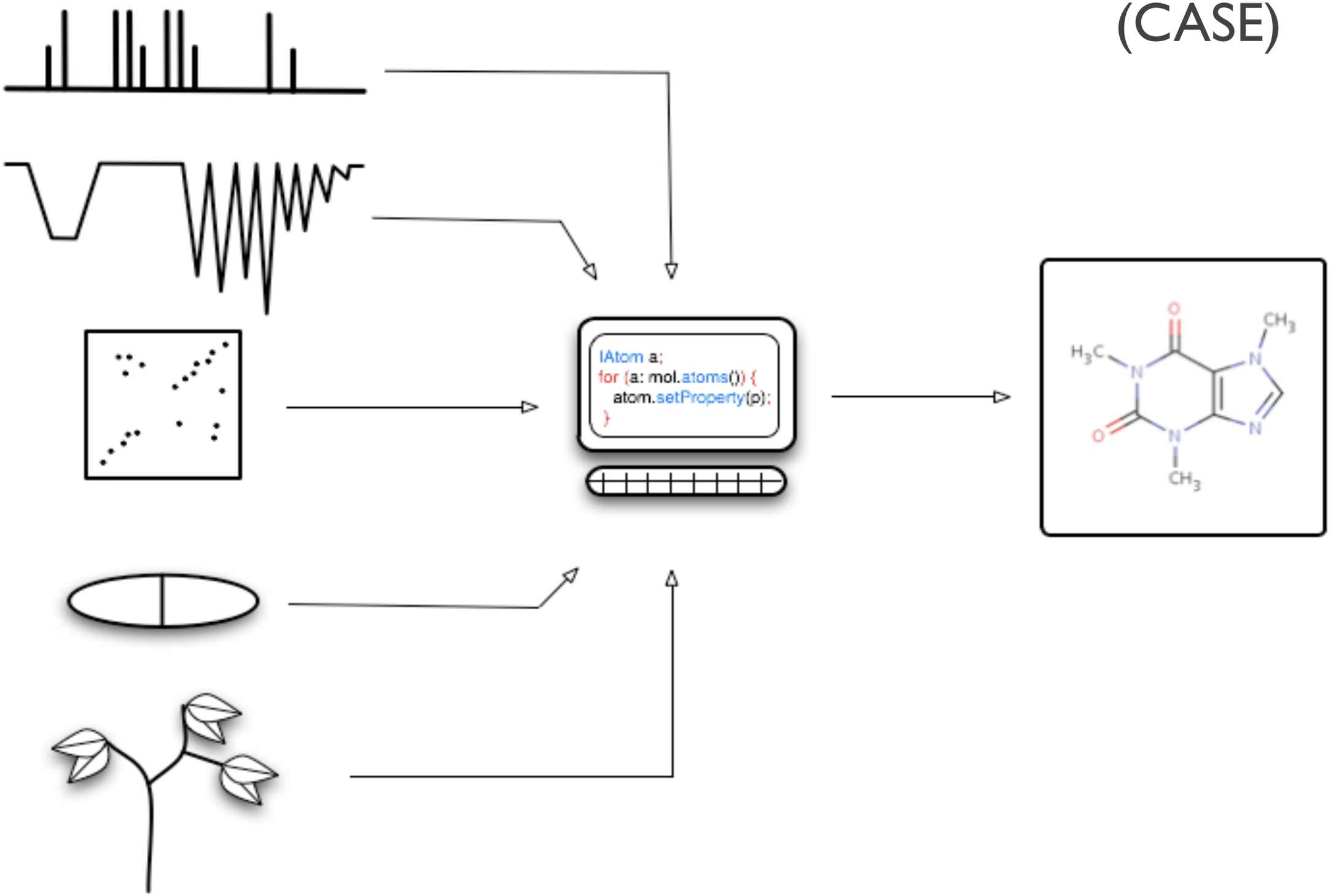


After evaluation of HMBC, we are looking at a molecular puzzle of pairs of carbon atoms that are either 1 or 2 bonds apart (but we don't know which is the case).



Computer-Assisted Structure Elucidation

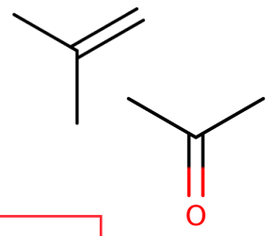
(CASE)



Computer-Assisted Structure Elucidation (CASE): Step by Step

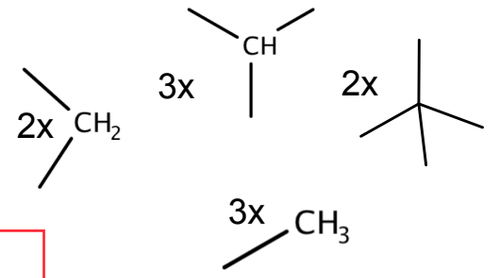
HR-MS, et al.

Gross Formula $C_{10}H_{16}$



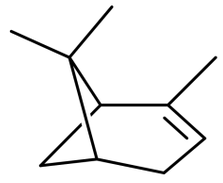
Typical ^{1}H and ^{13}C-chemical shifts

Functional Groups



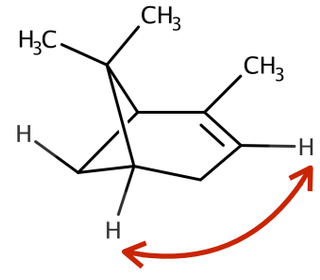
DEPT,
$^{1}J_{CH}$-Correlations

Structural Fragments



HH COSY, TOCSY
HMBC
1,1-ADEQUATE

Constitution



$^{3}J_{HH}$-Couplings
NOE-Diff. spectra
HH NOESY, HH ROESY

Relative Configuration

Complete Structure

Willighagen *et al. J Cheminform (2017) 9:33*
DOI 10.1186/s13321-017-0220-4

Journal of Cheminformatics

SOFTWARE

Open Access

The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching

Egon L. Willighagen[1*], John W. Mayfield[2], Jonathan Alvarsson[3], Arvid Berg[3], Lars Carlsson[4], Nina Jeliazkova[5], Stefan Kuhn[6], Tomáš Pluskal[7], Miquel Rojas-Chertó[8], Ola Spjuth[3], Gilleain Torrance[9], Chris T. Evelo[1], Rajarshi Guha[10] and Christoph Steinbeck[11]

Abstract

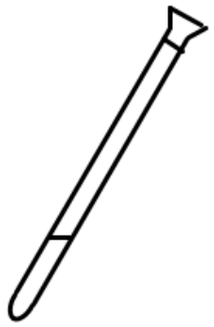
Background: The Chemistry Development Kit (CDK) is a widely used open source cheminformatics toolkit, providing data structures to represent chemical concepts along with methods to manipulate such structures and perform computations on them. The library implements a wide variety of cheminformatics algorithms ranging from chemical structure canonicalization to molecular descriptor calculations and pharmacophore perception. It is used in drug discovery, metabolomics, and toxicology. Over the last 10 years, the code base has grown significantly, however, resulting in many complex interdependencies among components and poor performance of many algorithms.

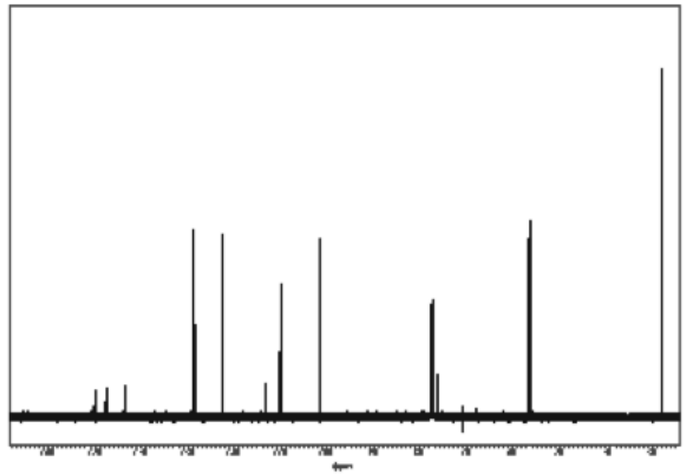
Results: We report improvements to the CDK v2.0 since the v1.2 release series, specifically addressing the increased functional complexity and poor performance. We first summarize the addition of new functionality, such atom typing and molecular formula handling, and improvement to existing functionality that has led to significantly better performance for substructure searching, molecular fingerprints, and rendering of molecules. Second, we outline how the CDK has evolved with respect to quality control and the approaches we have adopted to ensure stability, including a code review mechanism.

Conclusions: This paper highlights our continued efforts to provide a community driven, open source cheminformatics library, and shows that such collaborative projects can thrive over extended periods of time, resulting in a high-quality and performant library. By taking advantage of community support and contributions, we show that an open source cheminformatics project can act as a peer reviewed publishing platform for scientific computing software.

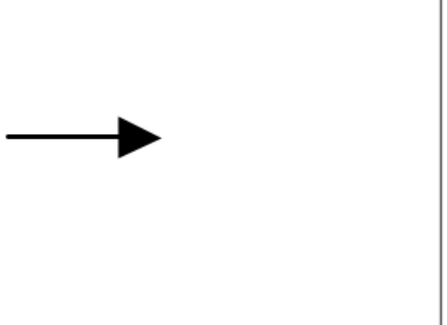
Keywords: Java, Cheminformatics, Bioinformatics, Metabolomics, Depiction

Structure Elucidation versus Identification/ Dereplication



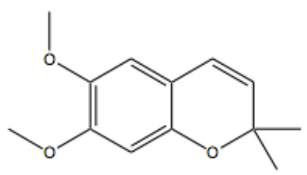
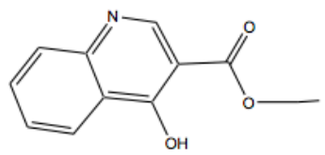
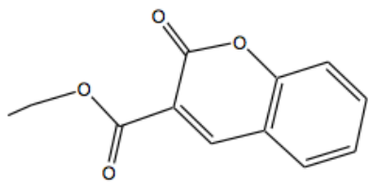
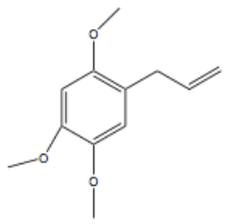
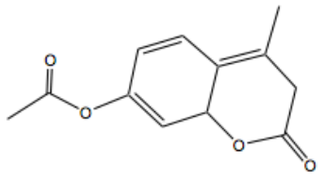
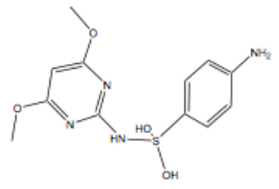


Shift [ppm]
149.99
147.58
143.46
128.64
122.38
113.44
110.01
101.39
76.42
56.90
56.32
28.08

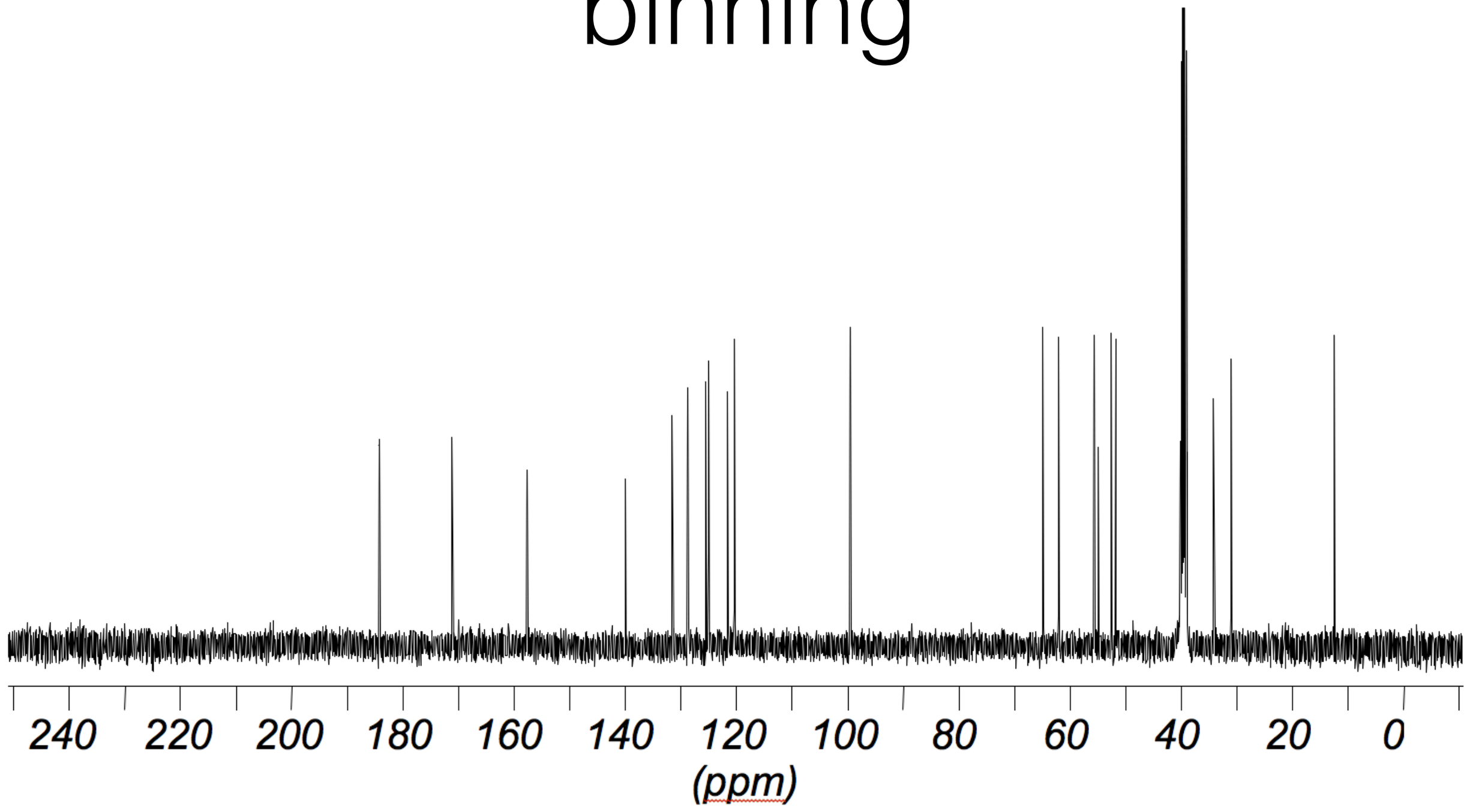


NMRShiftDB.org can do that for you

Steinbeck, C., & Kuhn, S. (2004). *Phytochemistry*, *65*(19), 2711–2717.

 Similarity :99.75%	 Similarity :58.97%
 Similarity :53.96%	 Similarity :52.67%
 Similarity :51.50%	 Similarity :51.00%

Fast structure searches via binning



0	0	0	0	0	0	0	1	0	1	0	0	1	1	1	1	0	1	0	0	0	1	1	0	1	1	0	1	0	0

40 years of CASE research

Applications of Artificial Intelligence for Chemical Inference. 22. Automatic Rule Formation in Mass Spectrometry by Means of the Meta-DENDRAL Program[1a]

B. G. Buchanan,* D. H. Smith, W. C. White, R. J. Gritter,[1b] E. A. Feigenbaum, J. Lederberg, and Carl Djerassi

Contribution from the Departments of Computer Science, Chemistry, and Genetics, Stanford University, Stanford, California 94305. *Received January 27, 1976*

Abstract: The DENDRAL computer program uses established rules of molecular fragmentation to help chemists solve complex structural problems from mass spectral data. This paper describes a computer program called Meta-DENDRAL, that can aid in the discovery of such rules from empirical data on known compounds. The program uses heuristic methods to search for common structural environments around those bonds that are found to fragment and abstracts plausible fragmentation rules. The program has been tested on the well-characterized, low-resolution mass spectra of aliphatic amines and the high-resolution mass spectra of estrogenic steroids. The program has also discovered new fragmentation rules for mono-, di-, and triketoandrostanes.

The DENDRAL computer program is designed to aid chemists with complex structure elucidation problems. One main part uses established molecular fragmentation rules to help chemists interpret mass spectra;[2] another main part generates lists of isomers that satisfy constraints derived from a variety of spectroscopic techniques.[3] Because the mass spectrometry rules used by the DENDRAL program have been culled from the literature, the program's growth depends upon manual examination of collections of spectra. But investigating the spectral data of new compound classes to determine frag-

J. Chem. Inf. Comput. Sci., Vol. 18, No. 4, 1978

CHEMICS-F: A Computer Program System for Structure Elucidation of Organic Compounds

SHIN-ICHI SASAKI,* HIDETSUGU ABE, YUJI HIROTA, YOSHIAKI ISHIDA, YOSHIHIRO KUDO, SHUKICHI OCHIAI, KEIJI SAITO, and TOHRU YAMASAKI

Miyagi University of Education, Aoba, Sendai 980 Japan

Received November 3, 1977

A computer program system CHEMICS-F for the structure elucidation of organic molecules containing C, H, and O is described in detail. CHEMICS-F involves the software used to analyze spectra, to convert the spectral information into "components" (defined substructures), and to construct a molecular structure based on the components by means of a newly developed method. Multiple structural formulas as final output are often constructed based on both desirable and undesirable components designated by the spectral information. To minimize the multiplicity of answers, a file handling procedure has been integrated in which the spectra of NMR, IR, and MS are stored in compressed and concentrated shapes capable of identification of compounds.

- Sesami
- Assemble
- Houdini

Fresenius Z Anal Chem (1982) 313:473–479

Fresenius Zeitschrift für
Analytische Chemie

Computer-Assisted Structure Elucidation

Morton E. Munk[1], Craig A. Shelley[2], Hugh B. Woodruff[3], and Mark O. Trulson[4]

[1] Department of Chemistry, Arizona State University, Tempe, Arizona 85287, USA
[2] Research Laboratories, B82, Kodak Park, Rochester, New York 14650, USA
[3] Merck, Sharp & Dohme Research Laboratories, P.O. Box 2000, Rahway, New Jersey 07065, USA
[4] Department of Chemistry, University of California, Berkeley, California 94720, USA

Strukturaufklärung mit Computer-Hilfe

Zusammenfassung. Ein in Entwicklung befindliches Computer-Modell („program CASE") wird beschrieben, das hauptsächlich folgende drei Aufgaben erfüllt: Zuordnung der chemischen und spektralen Eigenschaften einer Verbindung unbekannter Struktur zu ihren strukturellen Bedeutungen; Aufstellung aller Strukturen, die mit den gefundenen Eigenschaften im Einklang stehen; Einordnung dieser Strukturen je nach Übereinstimmung mit den vorhergesagten und gefundenen Parametern. Der derzeitige Entwicklungsstand des Programms wird erläutert und die Anwendung diskutiert.

Using intuition and experience! The solution may be a "procedure", perhaps resembling a set of rules.

Consider the problem of structure elucidation. A biochemist isolates an antibiotic from a fermentation beer, or an organic chemist detects a trace impurity in the commercial synthesis of a drug. The structure of the compound may be completely unknown or it is possible that it is related to a known compound and that a structure can be suggested. Of course, in the latter case confirmation is required. How does the chemist proceed? Generally, the chemical and spectral properties of the "unknown" are examined. Today there is a heavy emphasis on spectral properties because these data – the infrared spectrum, proton and carbon-13 nuclear mag-

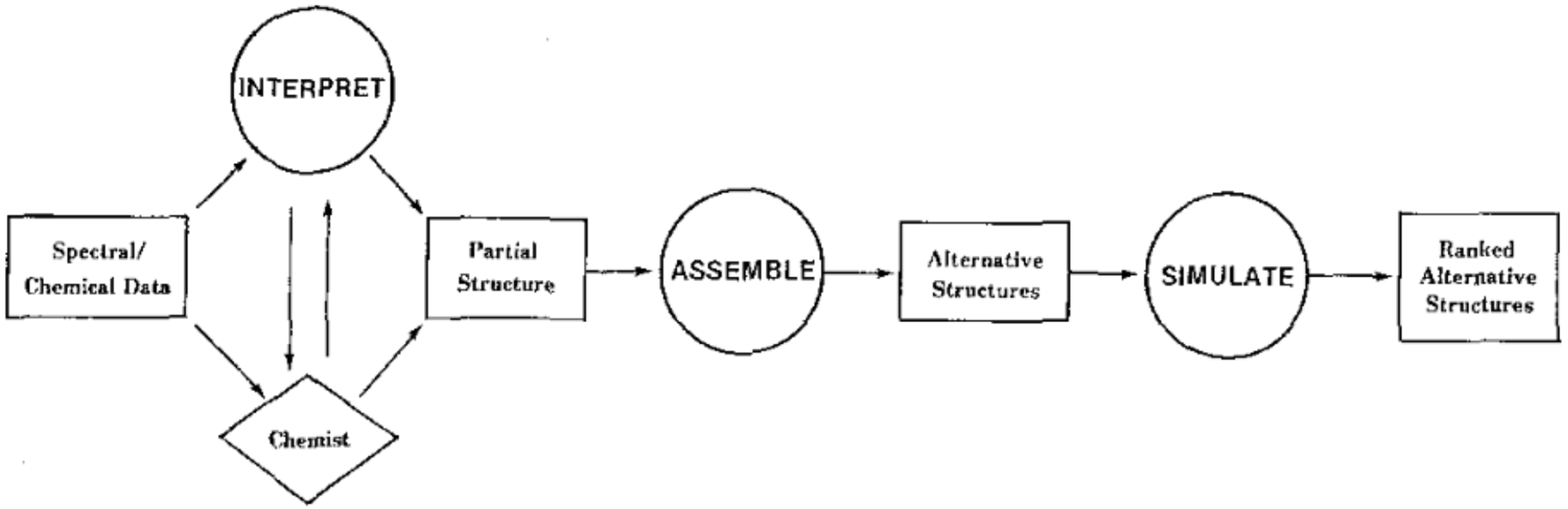


Fig. 1. Block diagram of C

Munk, M.E. et al., 1982. Computer-assisted structure elucidation. Fresenius' Zeitschrift für analytische Chemie, 313(6), pp.473–479.

Application of Expert System CISOC-SES to the Structure Elucidation of Complex Natural Products

Chen Peng,[†] Shengang Yuan,*[,†] Chongzhi Zheng,[†] Yongzheng Hui,[†] Houming Wu,[‡] and Kan Ma[‡]

Laboratory of Computer Chemistry and State Key Laboratory of Bioorganic and Natural Products Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, 354 Fenglin Lu, Shanghai 200032, People's Republic of China

Xiuwen Han

Dalian Institute of Chemical Physics, Chinese Academy of Sciences, P.O. Box 100, Dalian 116011, People's Republic of China

Received October 12, 1993®

This paper demonstrates the application of a newly developed computer-assisted structure elucidation system, CISOC-SES, to the structure determination of complex natural products. Discussed is the structure elucidation of four natural products, including one new compound, with up to 50 non-hydrogen atoms principally from their 2D NMR spectral data. The effectiveness of the novel approaches that exploit the direct and long-range distance constraints is analyzed in detail. The results show that the efficiency of structure elucidation can be substantially improved with the help of the system so that a manageable number of candidate structures, with the correct structure always included, can be obtained in a supportable period of CPU time.

LSD
(Logic for Structure Determination)

- LSD (Logic for Structure Determination
- Command-line driven
- Takes spectral constraints as input
- Generates lists of connection tables (molecules)
- Open Source
- Rocket-fast
- No early spectrum processing

Ab-Initio Structure Elucidation by 2D NMR

2D peak picking table

No.	^{13}C CPD	$^{1}J_{CH}$ (HMQC)	HMBC
1	148.24	-	4.0; 2.42; 2.28; 1.20
2	118.25	5.49	4.0; 2.28; 2.17
3	66.32	4.0	5.49
4	43.87	2.17	5.49; 4.0; 2.42; 1.32; 1.20; 0.86
5	41.40	2.14	5.49; 1.32; 0.86
6	38.32	-	2.42; 1.32; 1.20; 0.86
7	32.00	1.20/2.42	
8	31.52	2.28	2.42; 1.20
9	26.52	1.32	0.86; 2.17
10	21.46	0.86	1.20; 1.32

Table with heavy-atom relations

Internally generated by CASE program

	1	2	3	4	5	6	7	8	9	I I
1	X									
2		X								
3	O	O	X							
4		O	O	X						
5		O			X					
6						X				
7	O			O		O	X			
8	O	O						X		
9				O	O				X	
10				O	O	O			O	X

Steinbeck, C. Computer-Assisted Structure Elucidation. In *Handbook on Chemoinformatics.;* Gasteiger, J. Ed.; Wiley-VCH: Weinheim, 2003; Vol. 2; pp. 1378-1406.

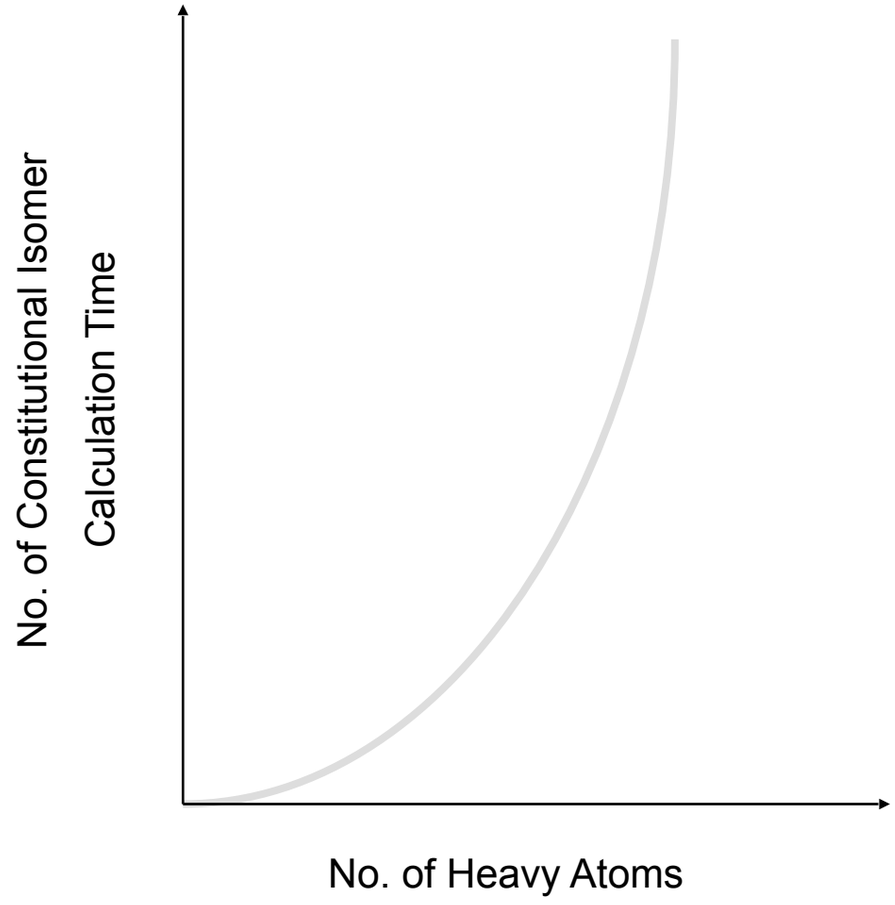
Limits to Growth

Limits to Growth

- Deterministic methods suffer from combinatorial explosion

Limits to Growth

- Deterministic methods suffer from combinatorial explosion

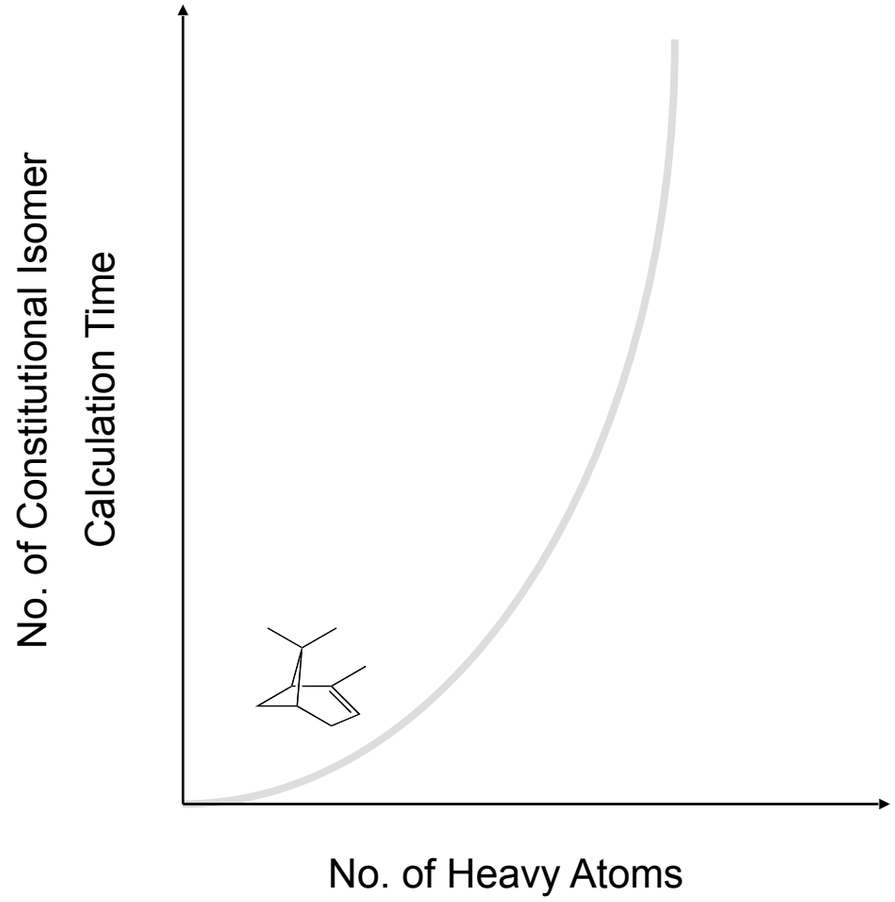


Limits to Growth

- Deterministic methods suffer from combinatorial explosion

$C_{10}H_{16}$ (10 Heavy Atoms)

24938 Constitutional Isomers



Limits to Growth

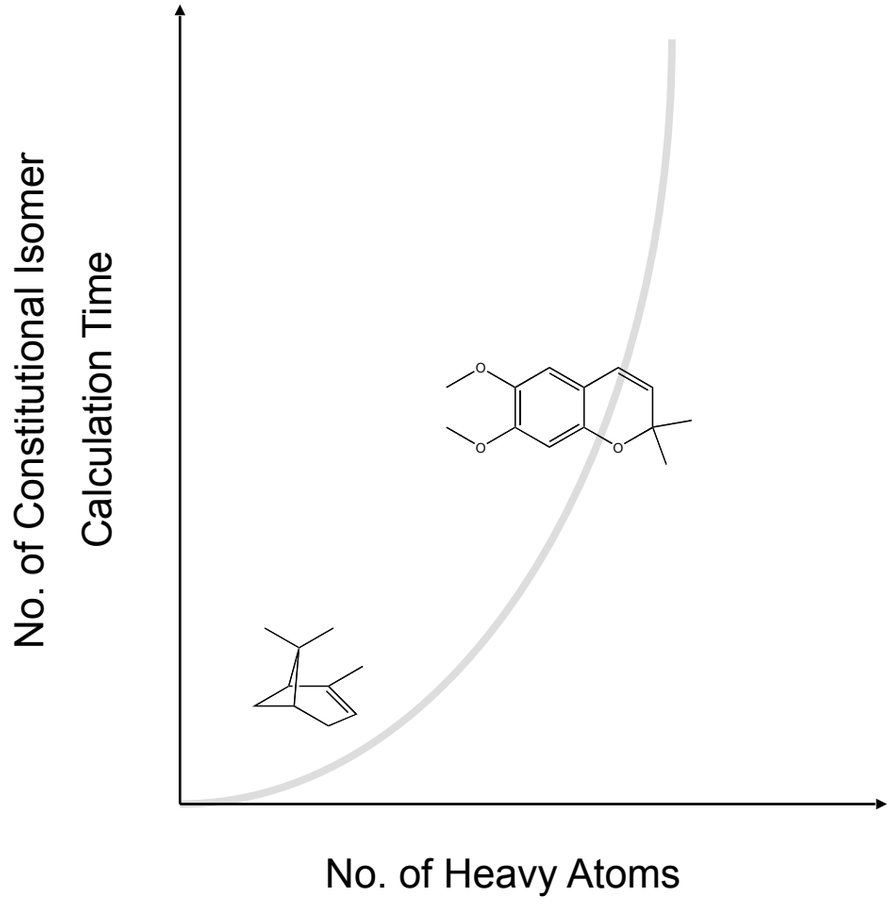
- Deterministic methods suffer from combinatorial explosion

$C_{10}H_{16}$ (10 Heavy Atoms)

24938 Constitutional Isomers

$C_{13}H_{16}O_3$ (16 Heavy Atoms)

> 2,000,000,000 Constitutional Isomers



Limits to Growth

- Deterministic methods suffer from combinatorial explosion

$C_{10}H_{16}$ (10 Heavy Atoms)

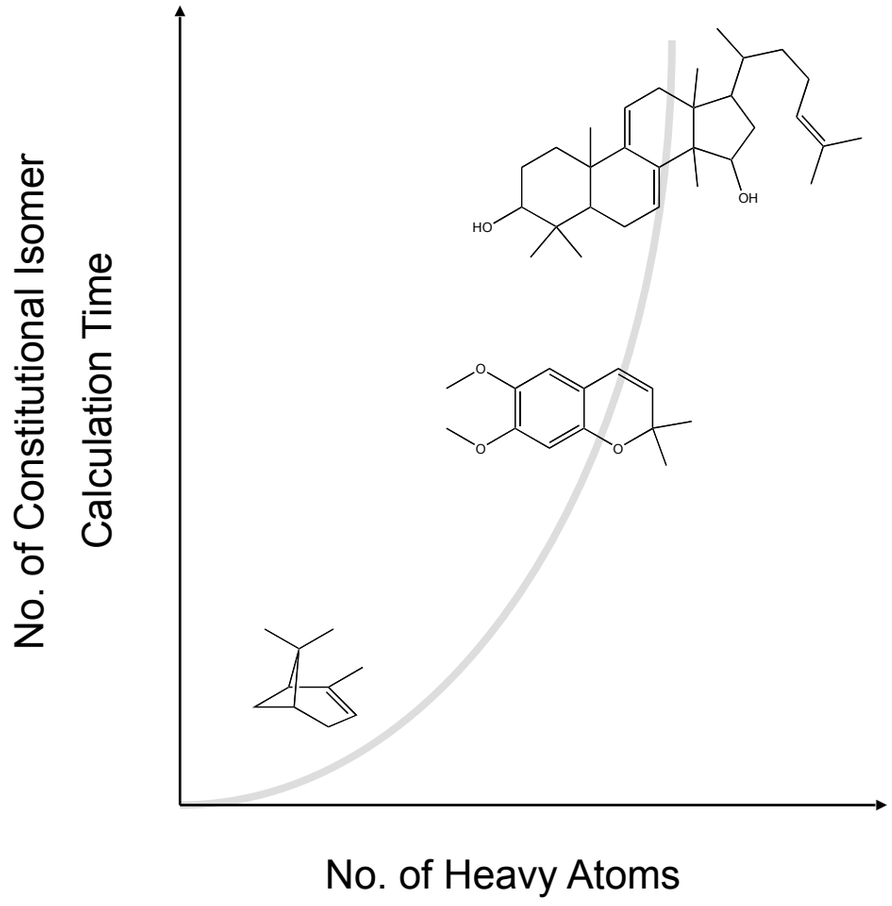
24938 Constitutional Isomers

$C_{13}H_{16}O_3$ (16 Heavy Atoms)

> 2,000,000,000 Constitutional Isomers

$C_{30}H_{48}O_2$ (32 Heavy Atoms)

>> 10^{12} Constitutional Isomers



Limits to Growth

• Deterministic methods suffer from combinatorial explosion

$C_{10}H_{16}$ (10 Heavy Atoms)

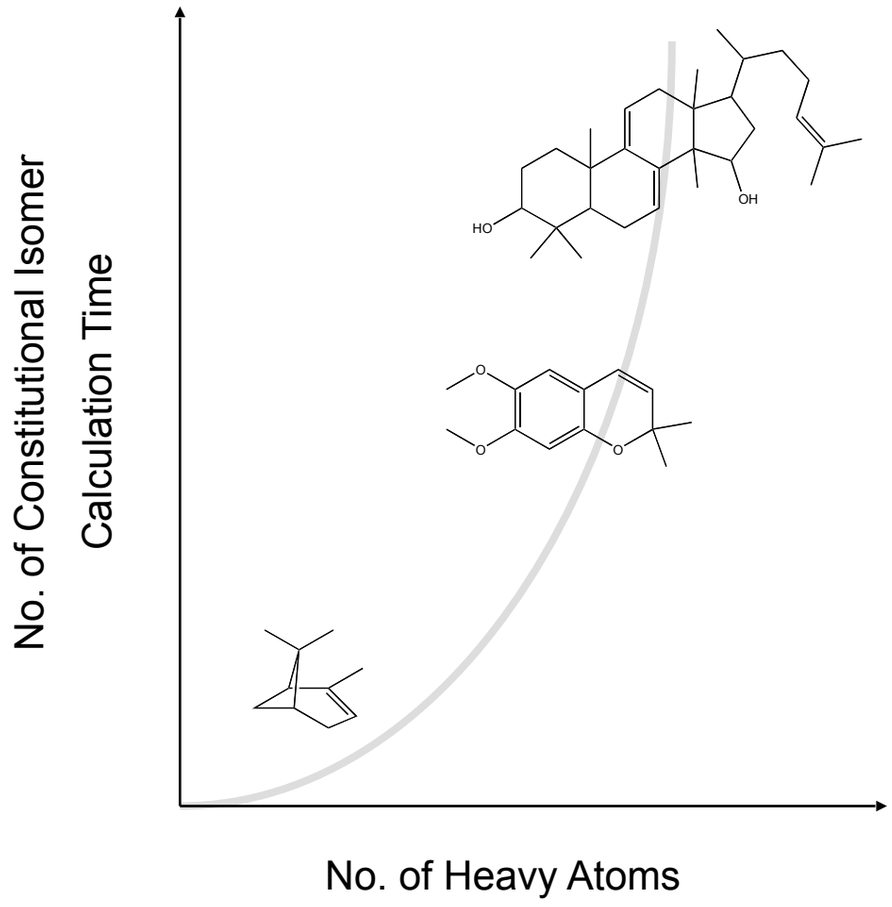
24938 Constitutional Isomers

$C_{13}H_{16}O_3$ (16 Heavy Atoms)

> 2,000,000,000 Constitutional Isomers

$C_{30}H_{48}O_2$ (32 Heavy Atoms)

>> 10^{12} Constitutional Isomers



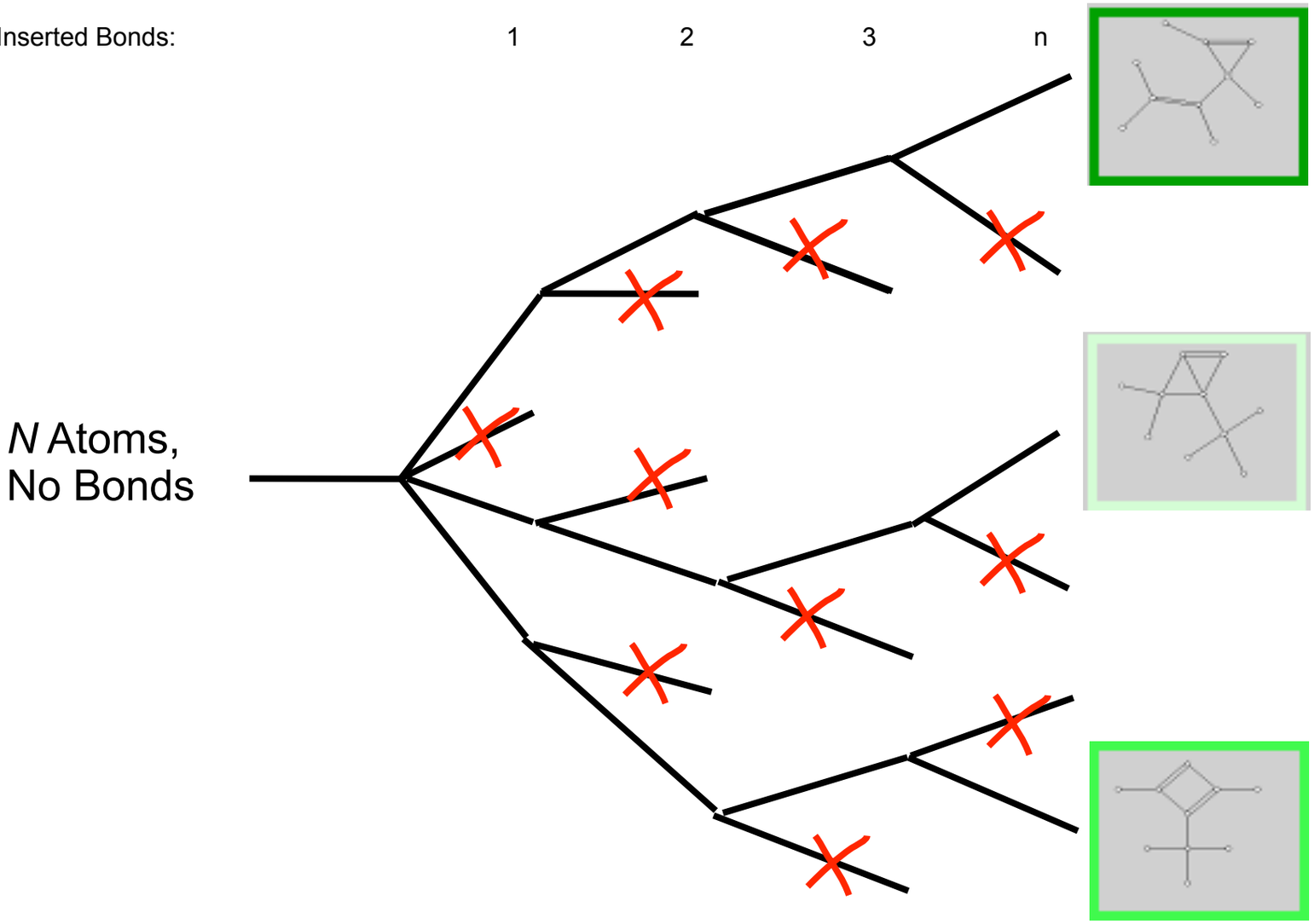
• Prospective use of spectroscopic input information may make them error-intolerant

Deterministic Structure Generators: The LUCY Method

- Prospective use of spectral information for building isomers
- Needs 1D ^{13}C, 2D HMQC, HMBC, HH COSY
- Example: Walk a decision tree while interpreting HMBC-Signals

	1	2	3	4	5	6	7	8	9	I I
1	X									
2		X								
3	O	O	X							
4		O	O	X						
5		O			X					
6						X				
7	O			O		O	X			
8	O	O						X		
9				O	O				X	
10				O	O	O			O	X

HMBC-derived relations between Heteroatoms



Steinbeck, C.; *Angewandte Chemie. International Ed. in English* **1996**, *35*, 1984-1986.

LSD Input Syntax: Basics

AcAlaOMe Example

Atom definitions

MULT 1 C 3 3
MULT 2 C 2 0
MULT 3 N 3 1
MULT 4 C 3 1
MULT 5 C 2 0
MULT 6 O 3 0
MULT 7 C 3 3
MULT 8 C 3 3
MULT 9 O 2 0
MULT 10 O 2 0

$^{1}J_{CH}$, redundant?

HMQC 1 1
HMQC 3 3
HMQC 4 4
HMQC 7 7
HMQC 8 8

Concatenate those blocks in a text file

$^{2/3}J_{CH}$, C-H-long-range correlations

HMBC 1 3
HMBC 2 1
HMBC 2 3
HMBC 3 1
HMBC 4 3
HMBC 4 8
HMBC 5 3
HMBC 5 4
HMBC 5 8
HMBC 5 7

http://eos.univ-reims.fr/LSD/index_ENG.html

LSD Usage

AcAlaOMe Example

Command prompt - don't type this in

```
$ lsd AcAlaOMe
```

```
$ outlsd 7 < acalaome.sol > acalaome.sdf
```

http://eos.univ-reims.fr/LSD/index_ENG.html

LSD Usage

AcAlaOMe Example

Command prompt - don't type this in

will produce acalaome`.sol`

(LSD specific output file)

```
$ lsd AcAlaOMe
```

```
$ outlsd 7 < acalaome.sol > acalaome.sdf
```

http://eos.univ-reims.fr/LSD/index_ENG.html

LSD Usage

AcAlaOMe Example

Command prompt - don't type this in

will produce acalaome`.sol`

(LSD specific output file)

```
$ lsd AcAlaOMe
```

```
$ outlsd 7 < acalaome.sol > acalaome.sdf
```

converts .sol into .sdf

(standard molecular file format)

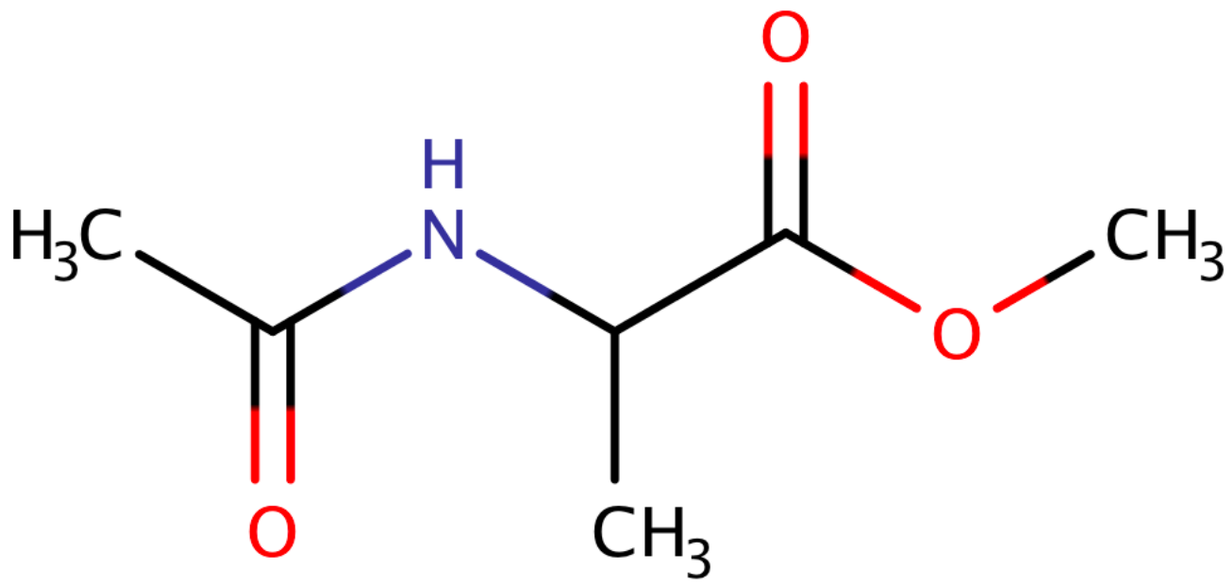
Use e.g. MarvinView to view .sdf

http://eos.univ-reims.fr/LSD/index_ENG.html

LSD Usage

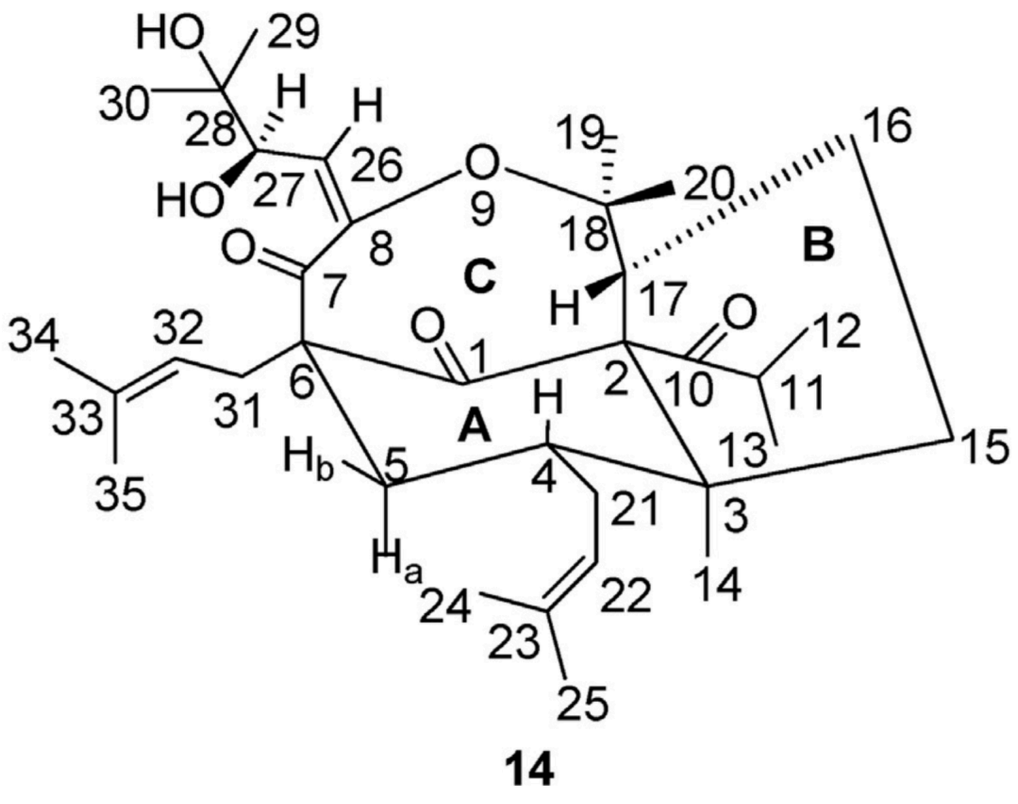
AcAlaOMe Example

MarvinView rendering of the one result in accordance with out input data from the previous slide



LSD CASE reveals one more possible solution

Liu et al., $C_{34}H_{52}O_6$, HRESI-MS (m/z 557.3833 [m+H])



14

No.	14		
	δ_C	δ_H muti (J/Hz)	HMBC
1	206.3		
2	74.5		
3	58.5		
4	39.2	1.70 m	C-2,C-3,C-5,C-6, C-14,C-15, C-22
5	35.6	(Ha)2.61dd(14.1,3.0) 1.21 m (Hb)1.21 m	C-1, C-3, C-4, C-6,C-7, C-21
6	61.3		
7	195.8		
8	150.6		
9			
10	217.0		
11	42.7	3.20 dp(12.8, 6.4)	C-10, C-12, C-13
12	22.6	1.38 d (6.5)	C-10, C-11, C-13
13	19.8	1.12 d (6.4)	C-10, C-11, C-12
14	17.1	0.90 s	C-2, C-3, C-4, C-15
15	35.3	1.87ddd(12.6,9.7, 5.1), 1.42 m	C-2,C-3,C-4, C-14,C-16,C-17
16	25.7	2.32 m	C-3,C-15,C-17,C-18
17	55.3	2.88 t(10.0)	C-1,C-2,C-3,C-10,C-11, C-16,C-18,C-19,C-20
18	86.1		
19	24.9	1.19 s	C-17,C-18,C-20
20	24.2	1.15 s	C-17,C-18,C-19
21	29.4	2.02dd(12.8,5.5), 1.66 m	C-3,C-4,C-5, C-22, C-23
22	122.6	5.16 t(6.4)	C-4,C-21, C-23, C-24,C-25
23	133.4		
24	25.8	1.76 s	C-22, C-23, C-25
25	18.0	1.60 s	C-22, C-23, C-24
26	125.1	5.92 d(6.8)	C-7, C-8, C-18, C-27,C-28
27	74.0	4.19 d(6.8)	C-7,C-8,C-28, C-26,C-29,C-30
28	72.4		
29	26.6	1.19 s	C-27,C-28,C-30
30	25.8	1.15 s	C-27,C-28,C-29
31	34.1	2.35 m, 2.27 m	C-1,C-5,C-6, C-7,C-32,C-33
32	117.6	4.96 t(7.2)	C-6,C-31,C-33, C-34
33	135.2		
34	26.0	1.69 s	C-32,C-33,C-35
35	18.1	1.55 s	C-32,C-33,C-34

Liu, R., Su, Y., Yang, J., & Wang, A. (2017). Polyprenylated acylphloroglucinols from Hypericum scabrum. Phytochemistry, 142, 38–50.

LSD Input Syntax: Advanced

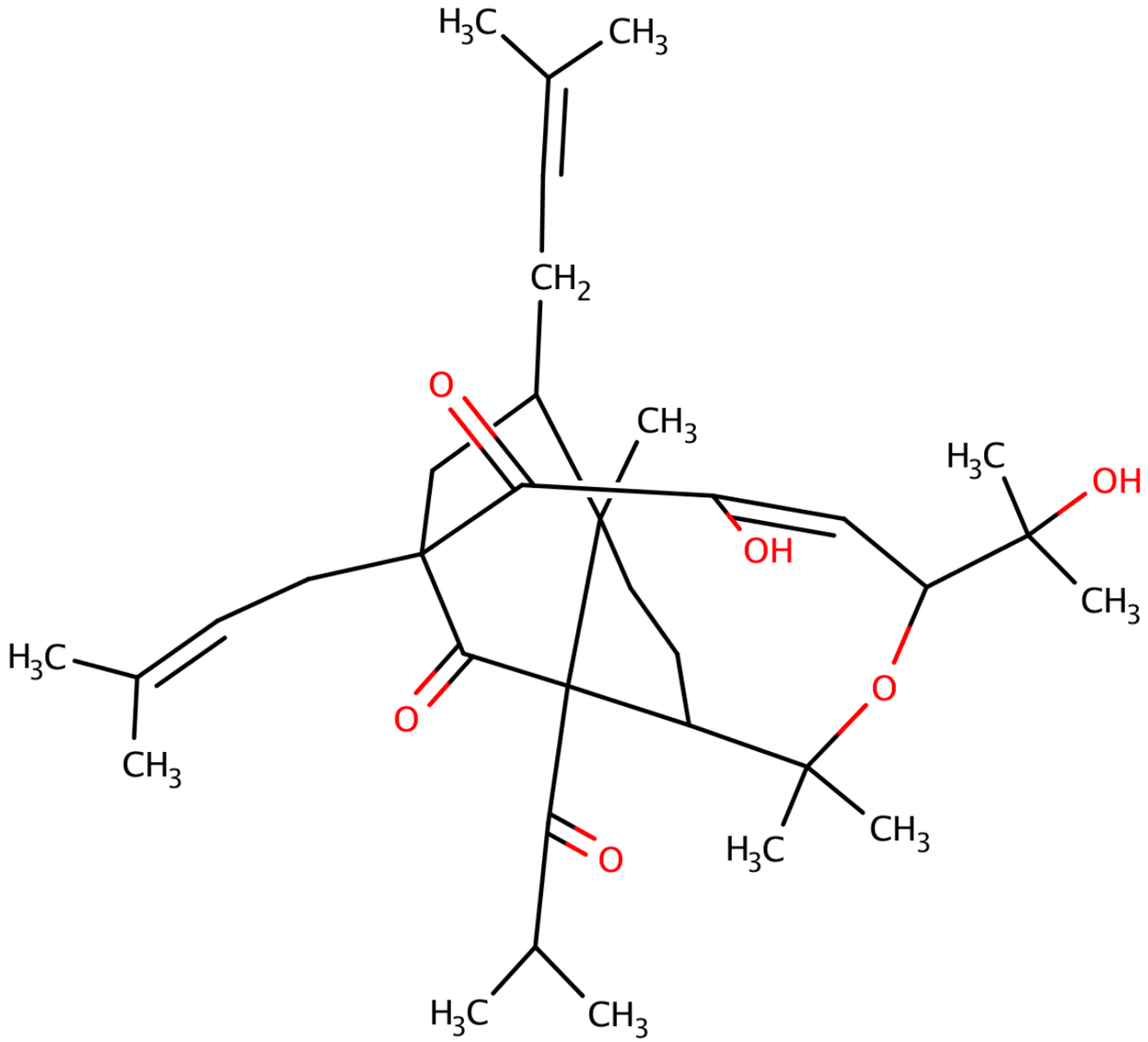
Liu et al., $C_{34}H_{52}O_6$, HRESI-MS (m/z 557.3833 [m+H])

```
ELIM 3 4

MULT 1 C 2 0
MULT 2 C 3 0
[… 35 more omitted …]

; known carbonyls
BOND 1 36
BOND 7 37
BOND 10 39

CARB L1 ; define list L1 containing all carbons
HETE L2
LIST L3 2 3 6 8 17 18 27 28
LIST L4 9 38 40
```

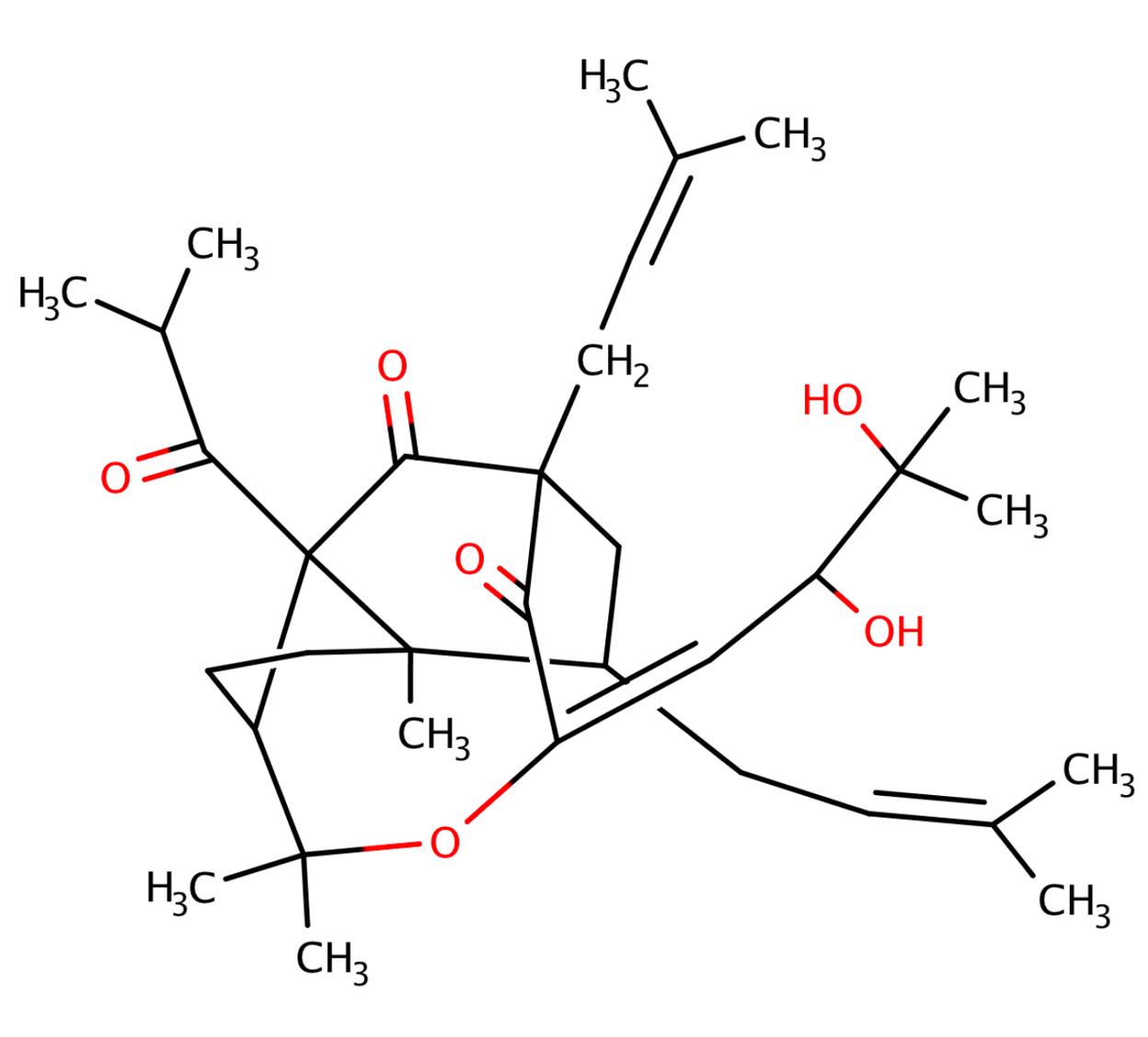
```
PROP L1 1 L2 - ; Every carbon atom
               ; can carry one or less
               ; hetero-atoms, but not
               ; two

PROP L4 0 L3 ; Every oxygen which is
             ; not an sp2 O has a
             ; parter from L3 (based
             ; on a conservative
             ; chemical shift
             ; inspection)

COSY 4 5
[… 4 more omitted …]

HMQC 4 4
[… more omitted …]

HMBC 2 4
[… 98 more omitted …]
```

http://eos.univ-reims.fr/LSD/index_ENG.html

Liu et al., $C_{34}H_{52}O_6$, HRESI-MS (m/z 557.3833 [m+H])

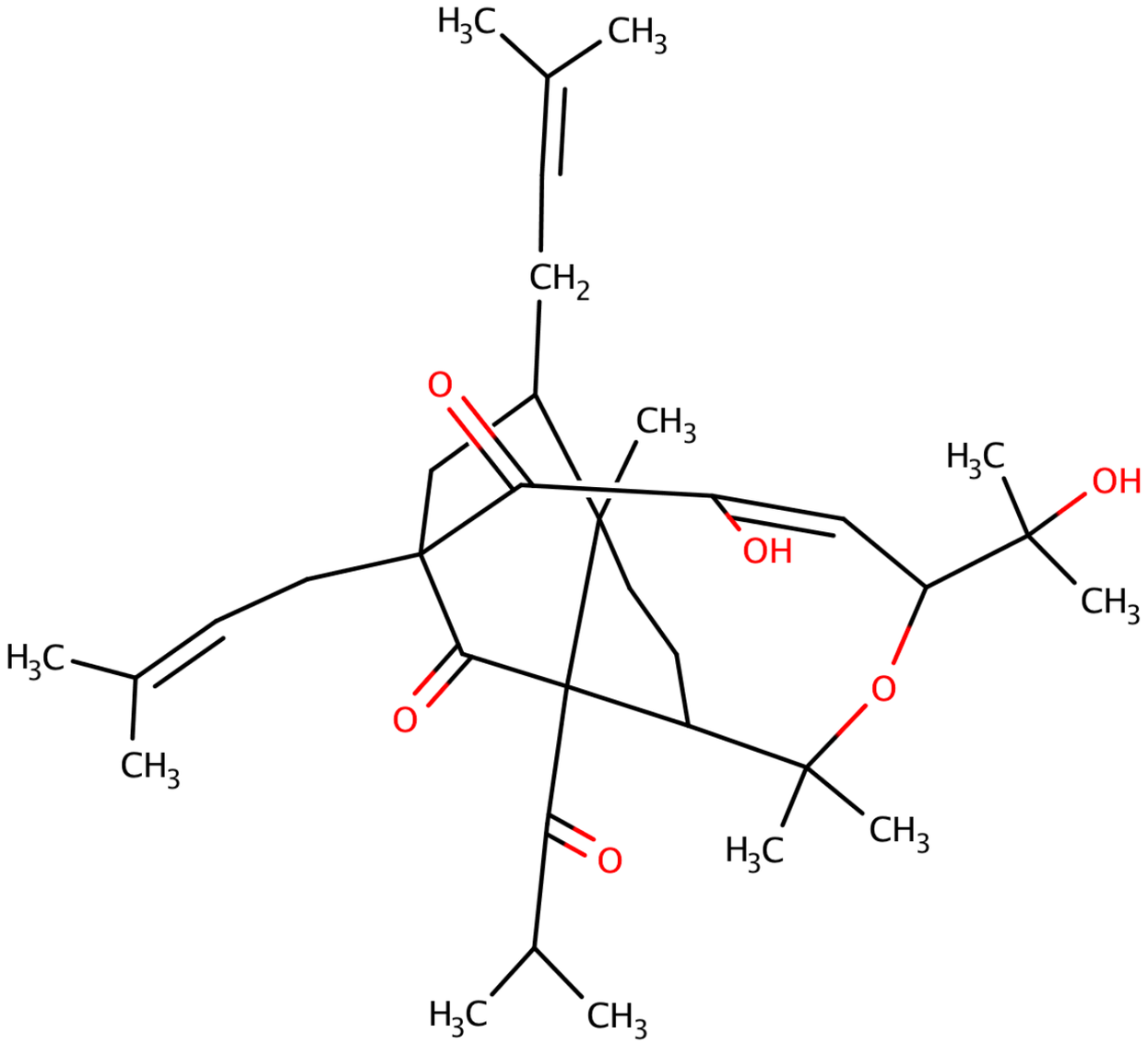


InChIKey=ANWFPAAUCGPEBV-MOHJPFBDNA-N

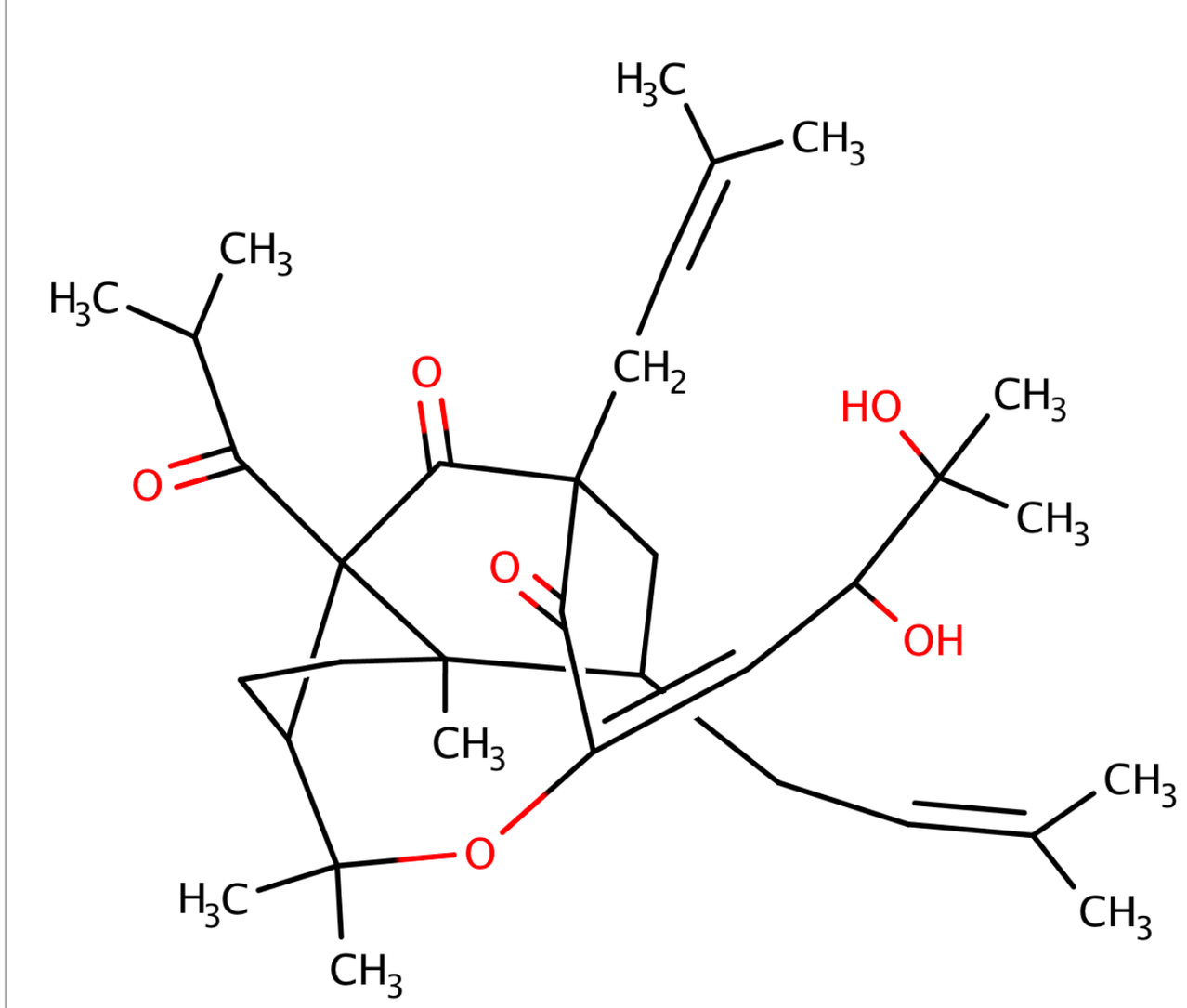


InChIKey=YAJAXOAXTCGOQA-HKOYGPOVNA-N

Liu et al., $C_{34}H_{52}O_6$, HRESI-MS (m/z 557.3833 [m+H])



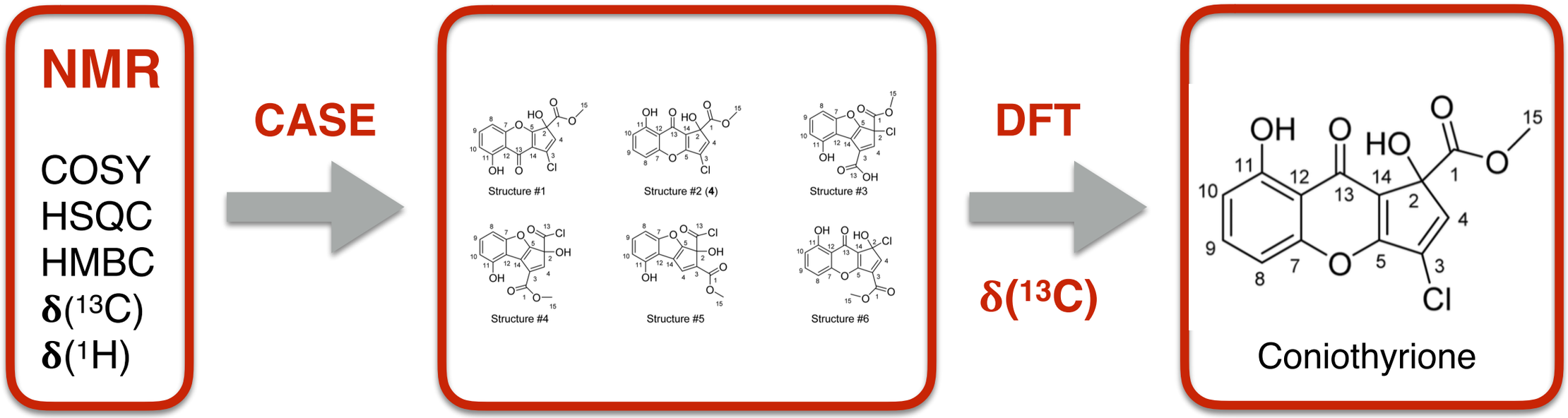
InChIKey=ANWFPAAUCGPEBV-MOHJPFBDNA-N



InChIKey=YAJAXOAXTCGOQA-HKOYGPOVNA-N

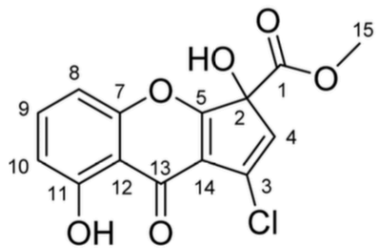
InChIKey of published compound #14 = YAJAXOAXTCGOQA-HKOYGPOVNA-N

Quantum Chemistry Calculations for Validation

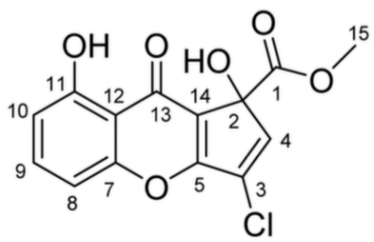


- Commercial Software (Gaussian 16, Spartan)
- Open Software (NWChem)
- Tricky setup, tricky analysis
- Potentially long calculation time (> day on your PC)

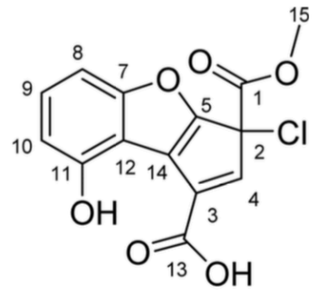
1. Buevich, A. V.; Elyashberg, M. E. Synergistic Combination of CASE Algorithms and DFT Chemical Shift Predictions: A Powerful Approach for Structure Elucidation, Verification, and Revision. J. Nat. Prod 2016, 79, 3105–3116.



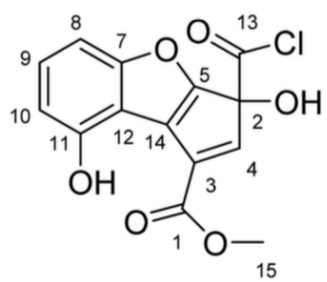
Structure #1



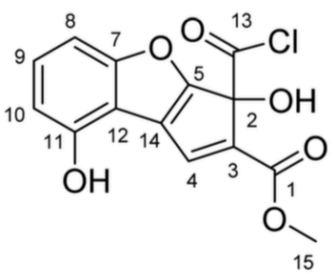
Structure #2 (**4**)



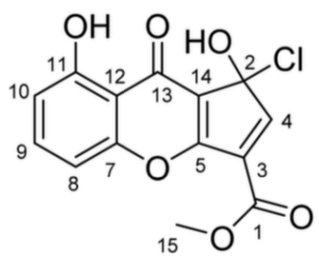
Structure #3



Structure #4



Structure #5



Structure #6

label	exptl	structure #1	structure #2 (**4**)	structure #3	structure #4	structure #5	structure #6
	δC	δC_{calc}	δC_{calc}	δC_{calc}	δC_{calc}	δC_{calc}	δC_{calc}
C1	**168.7**	171.3	172.0	165.9	165.6	164.0	160.4
C2	**79.8**	81.6	79.6	69.6	85.1	84.4	102.9
C3	**127.2**	142.7	135.6	146.1	147.1	140.9	156.3
C4	**143.1**	125.4	144.9	134.1	135.8	136.1	125.7
C5	**164.5**	171.0	163.8	160.8	155.7	163.8	164.1
C7	**155.7**	154.6	154.7	158.8	160.3	159.6	154.7
C8	**108.1**	105.3	105.5	105.3	101.5	103.8	105.7
C9	**135.8**	134.8	134.3	125.3	126.4	126.7	134.1
C10	**112.7**	111.9	112.0	106.5	108.9	109.1	111.8
C11	**160.6**	160.5	160.5	145.3	150.2	148.9	160.8
C12	**110.7**	109.8	110.3	110.2	108.8	108.6	110.1
C13	**176.0**	173.9	172.9	159.1	173.1	180.5	172.3
C14	**120.8**	117.2	120.6	121.1	124.7	122.7	117.3
C15	**52.9**	54.1	53.7	53.8	53.2	51.7	52.0
RMSD, ppm		6.75	2.76	9.46	7.86	6.40	11.29
max_δ, ppm		17.71	8.39	18.94	19.87	13.67	29.07

1. Buevich, A. V.; Elyashberg, M. E. Synergistic Combination of CASE Algorithms and DFT Chemical Shift Predictions: A Powerful Approach for Structure Elucidation, Verification, and Revision. J. Nat. Prod 2016, 79, 3105–3116.

Stochastic CASE

Stochastic Search Methods

Algorithms known to tackle large search spaces:

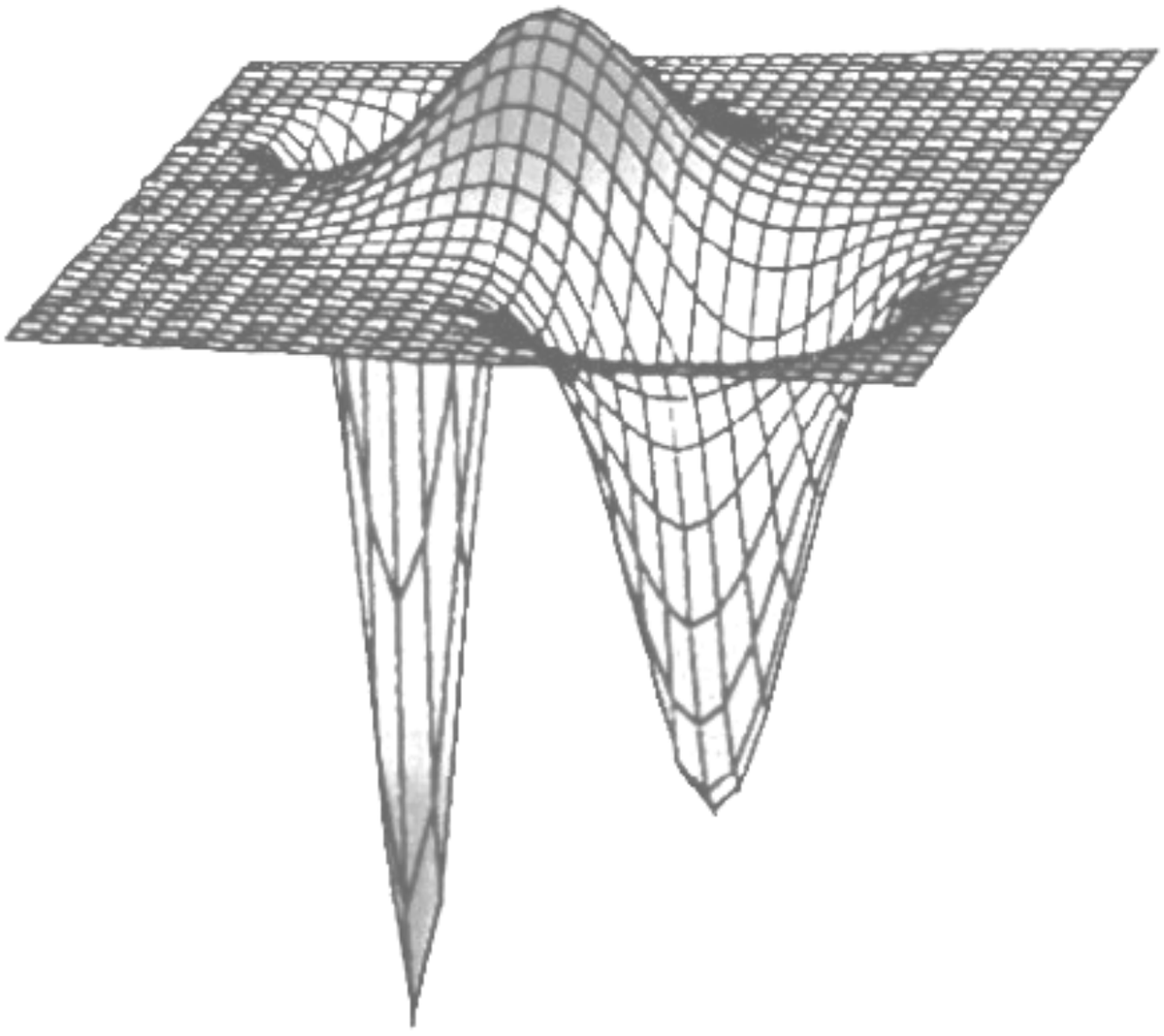
- *Simulated Annealing*
    - Traveling Salesman Problem
    - Finding the solution structure of large biomolecules
    - Integrated Circuits Layout
    - Robotic Path Planning

- *Genetic Algorithms*
    - Protein Folding
    - Immune System Simulation
    - Computer-Aided Design

- Quite a number of other options ...

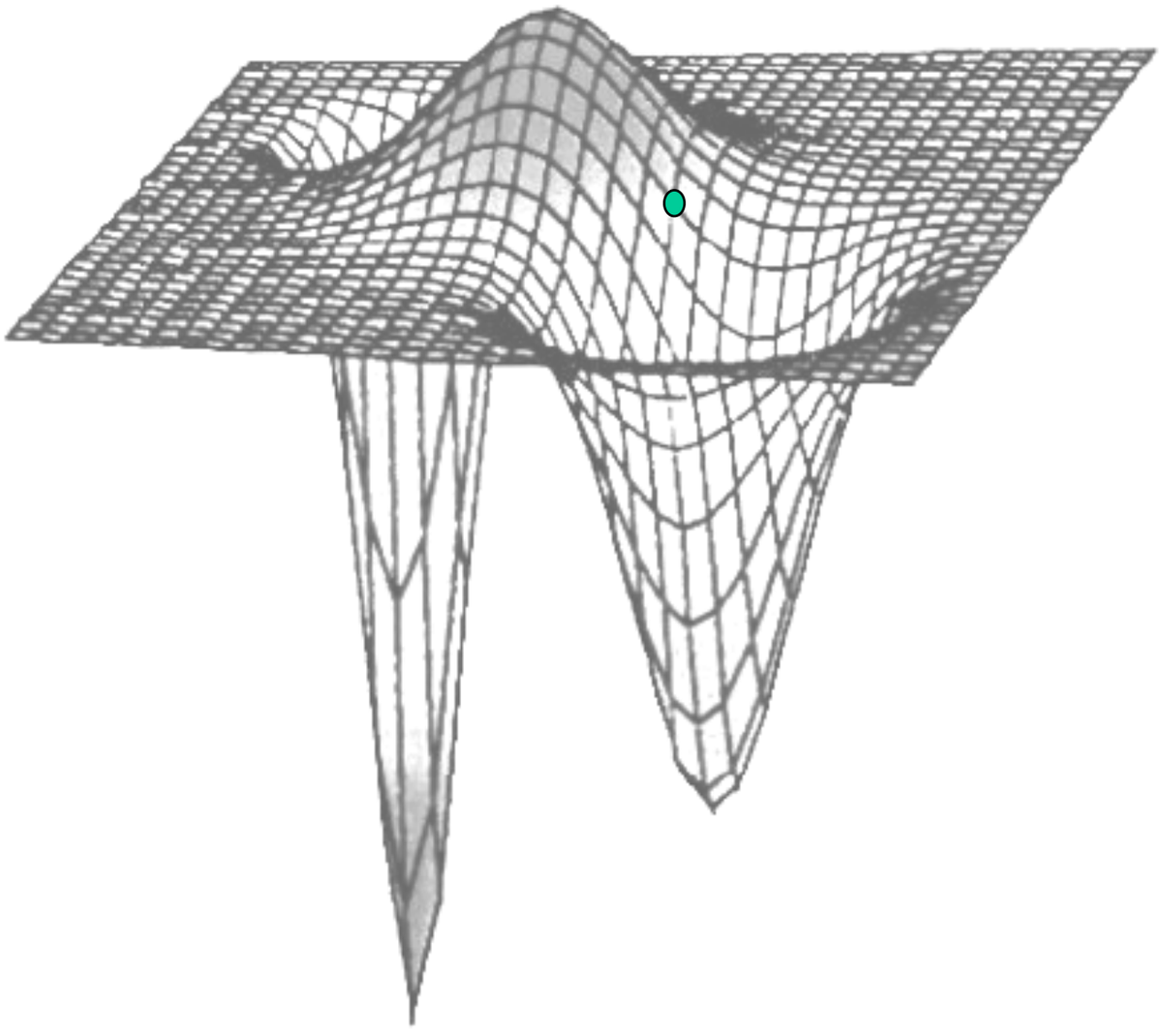
Simulated Annealing

Guided Walk in Constitution Space



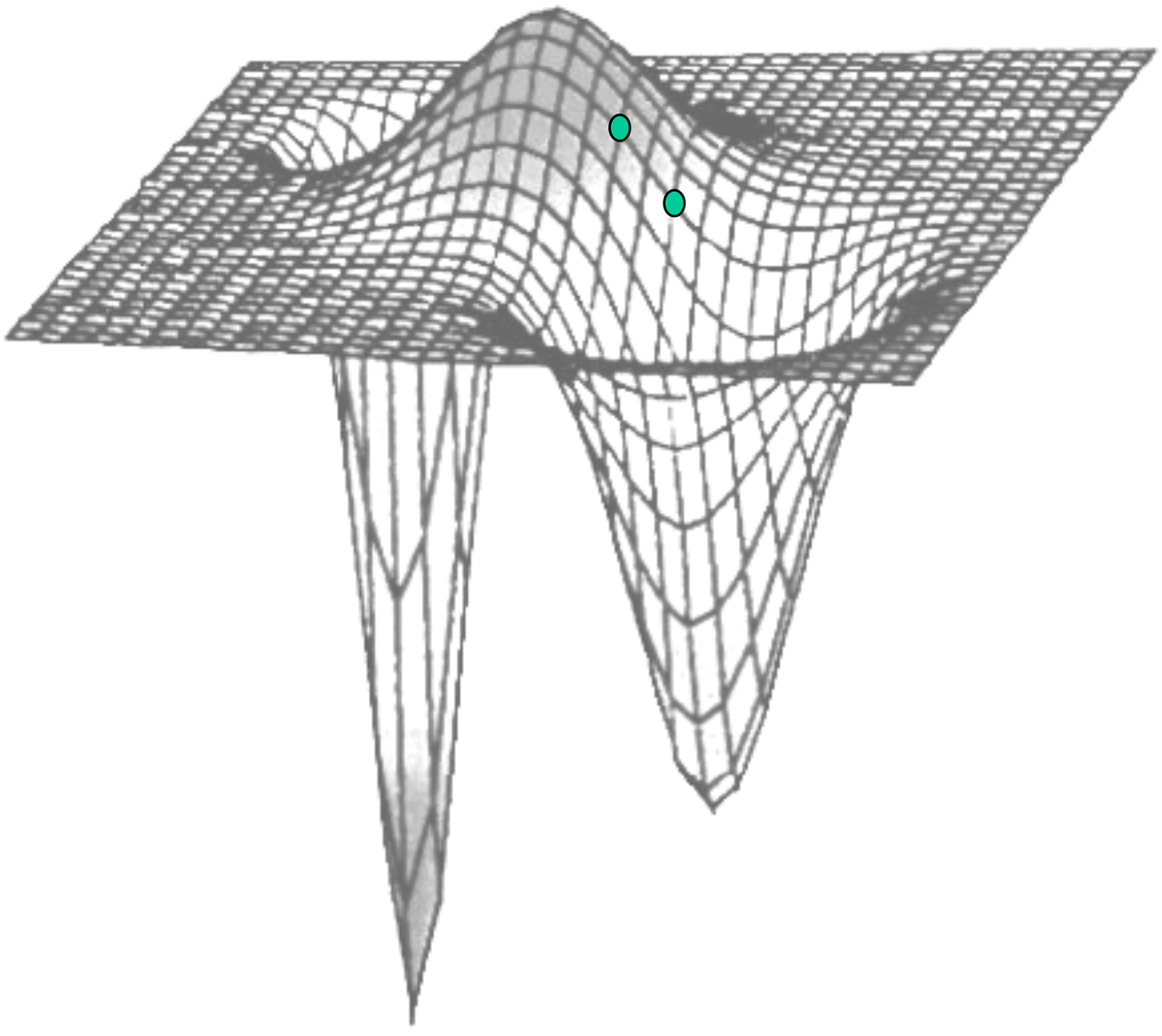
Simulated Annealing

Guided Walk in Constitution Space



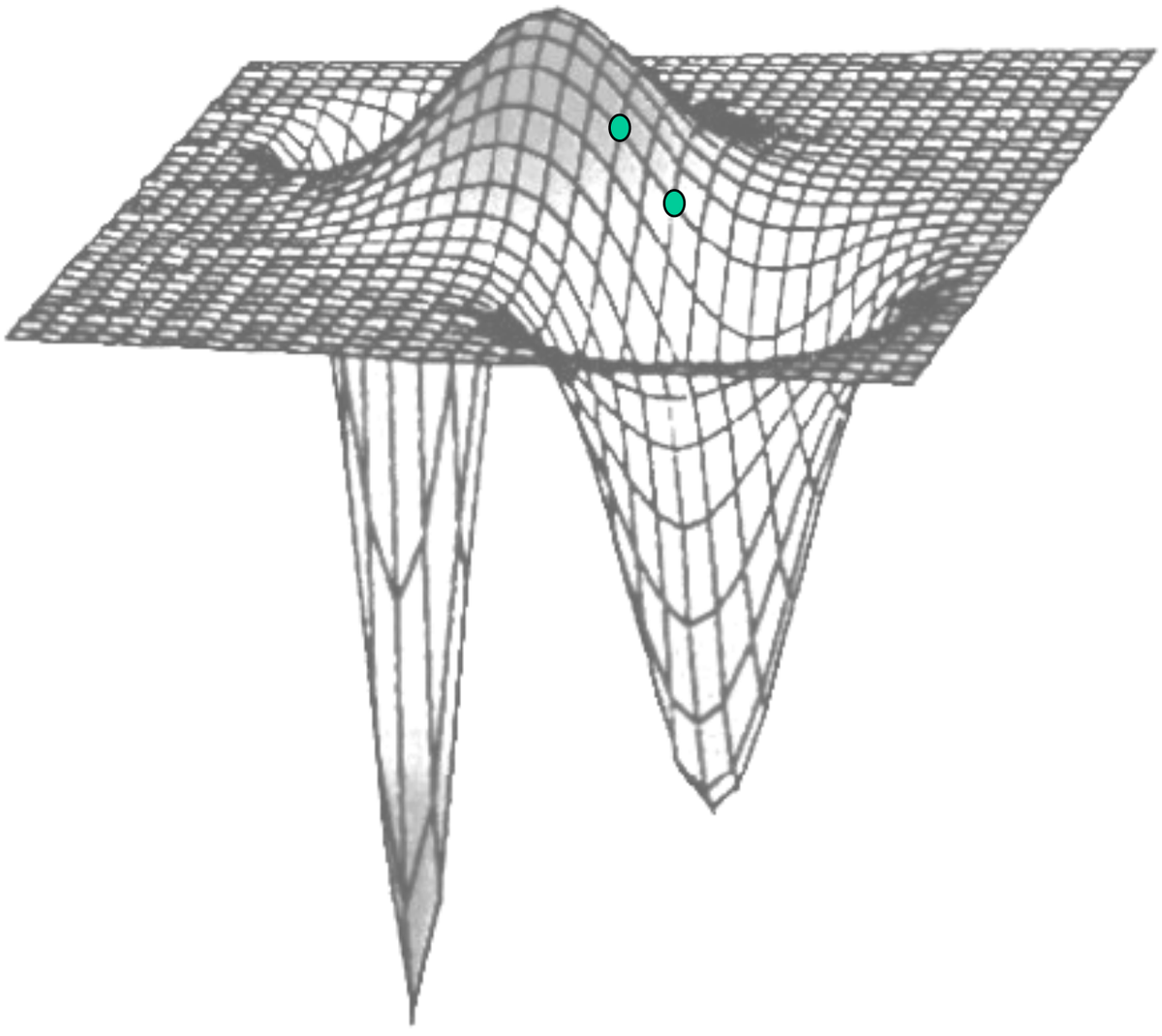
Simulated Annealing

Guided Walk in Constitution Space



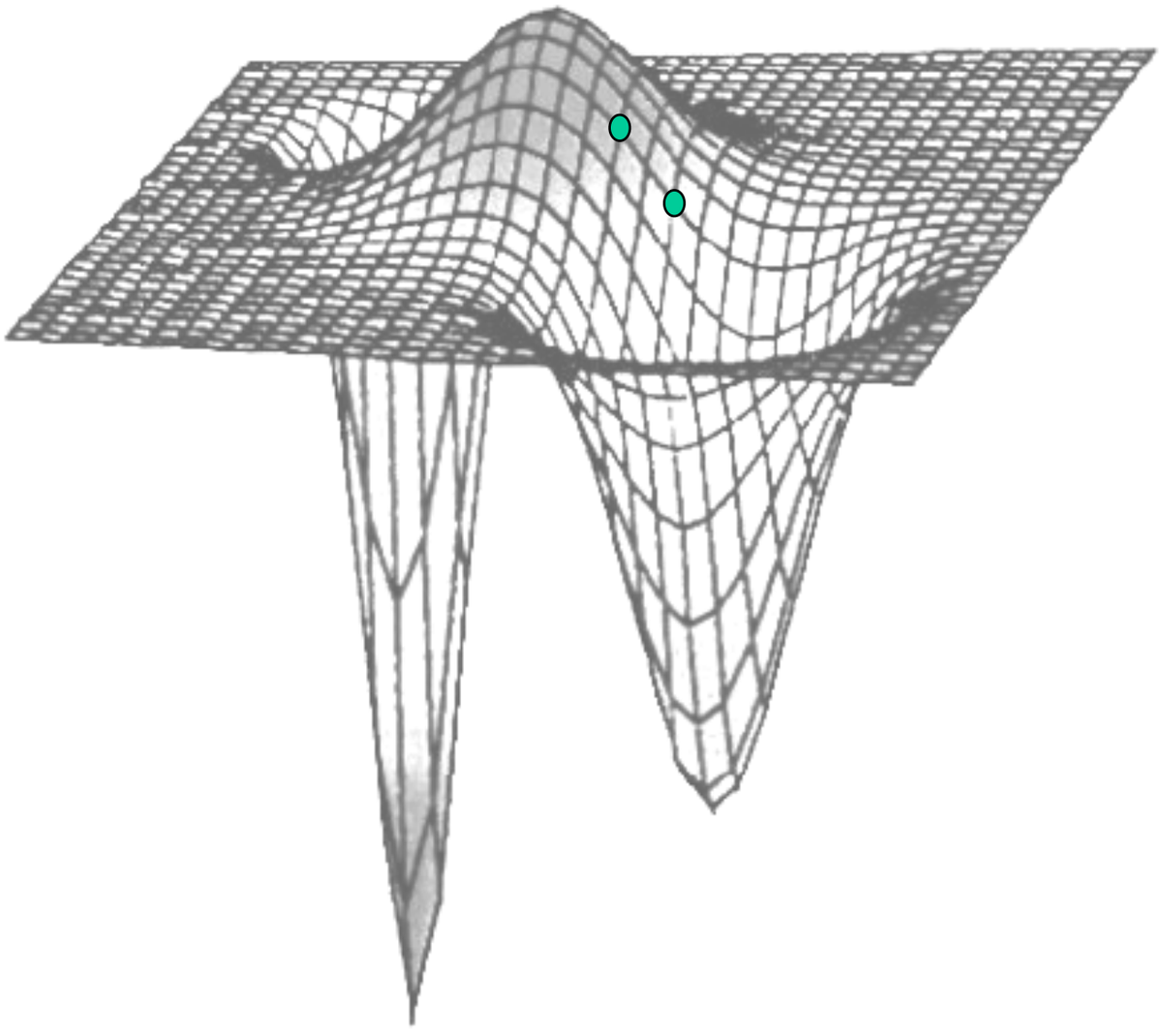
Simulated Annealing

Guided Walk in Constitution Space



Simulated Annealing

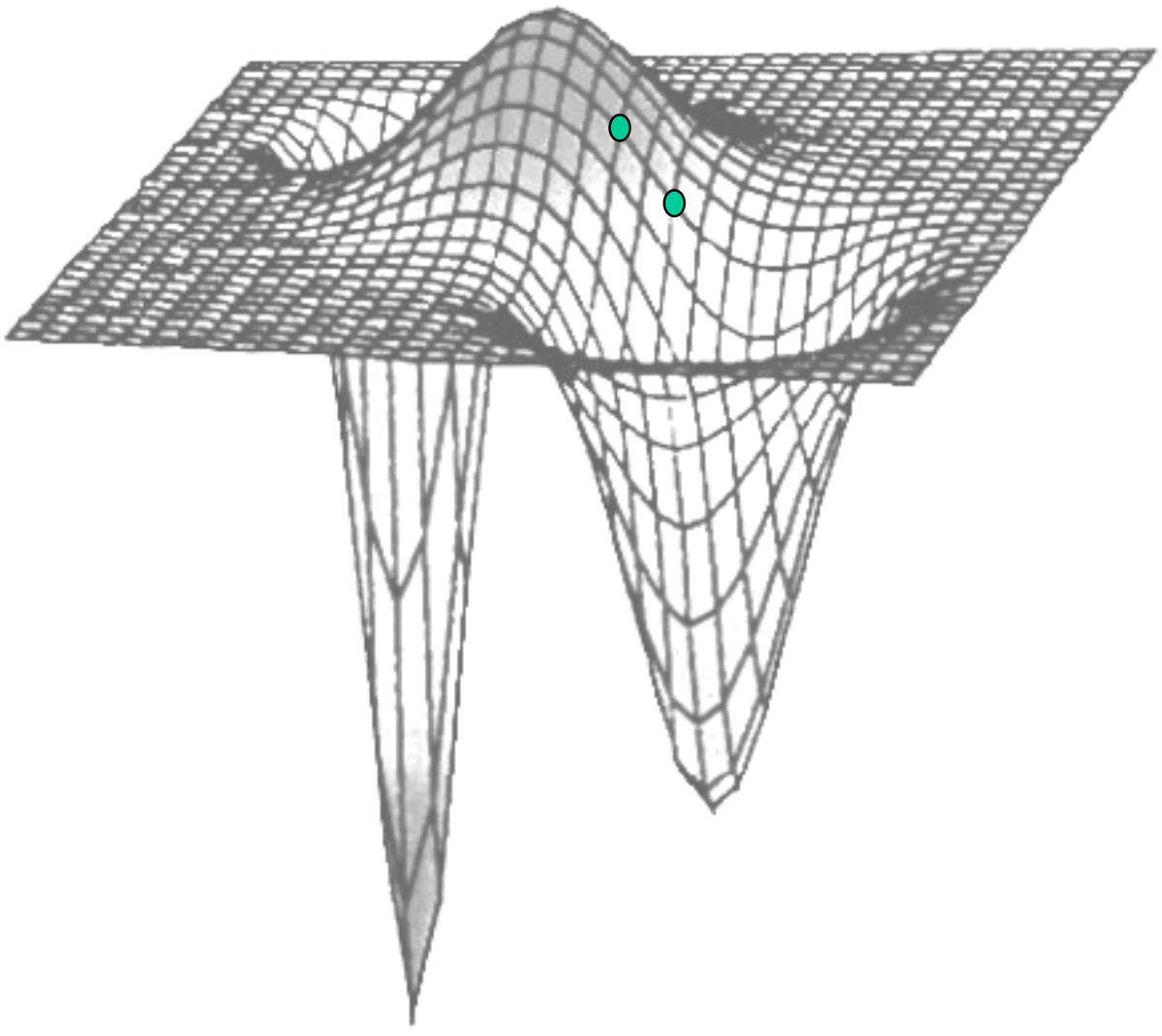
Guided Walk in Constitution Space



Neighbors in constitution space

Simulated Annealing

Guided Walk in Constitution Space

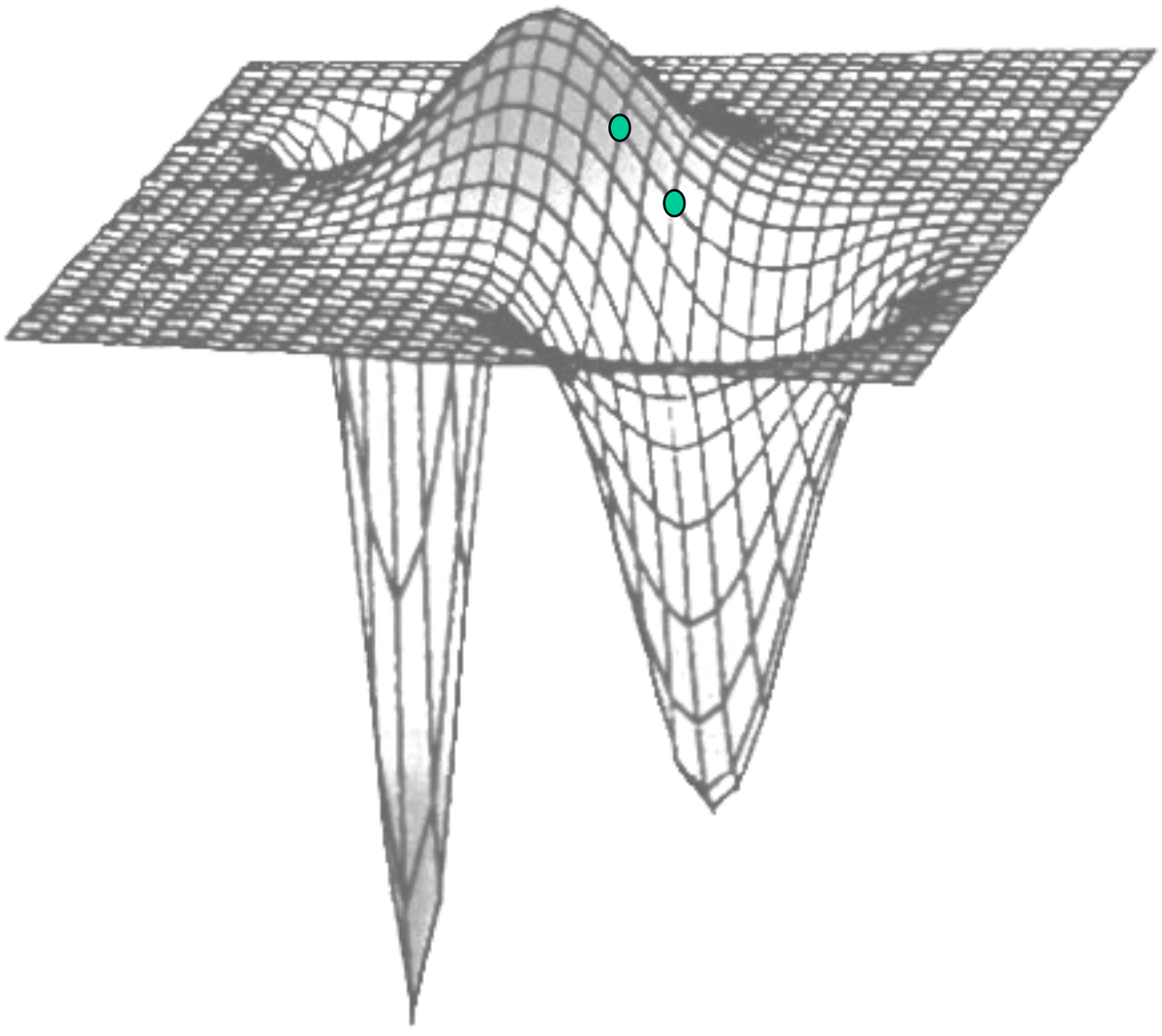


Neighbors in constitution space

- are in close chemical distance to each other

Simulated Annealing

Guided Walk in Constitution Space

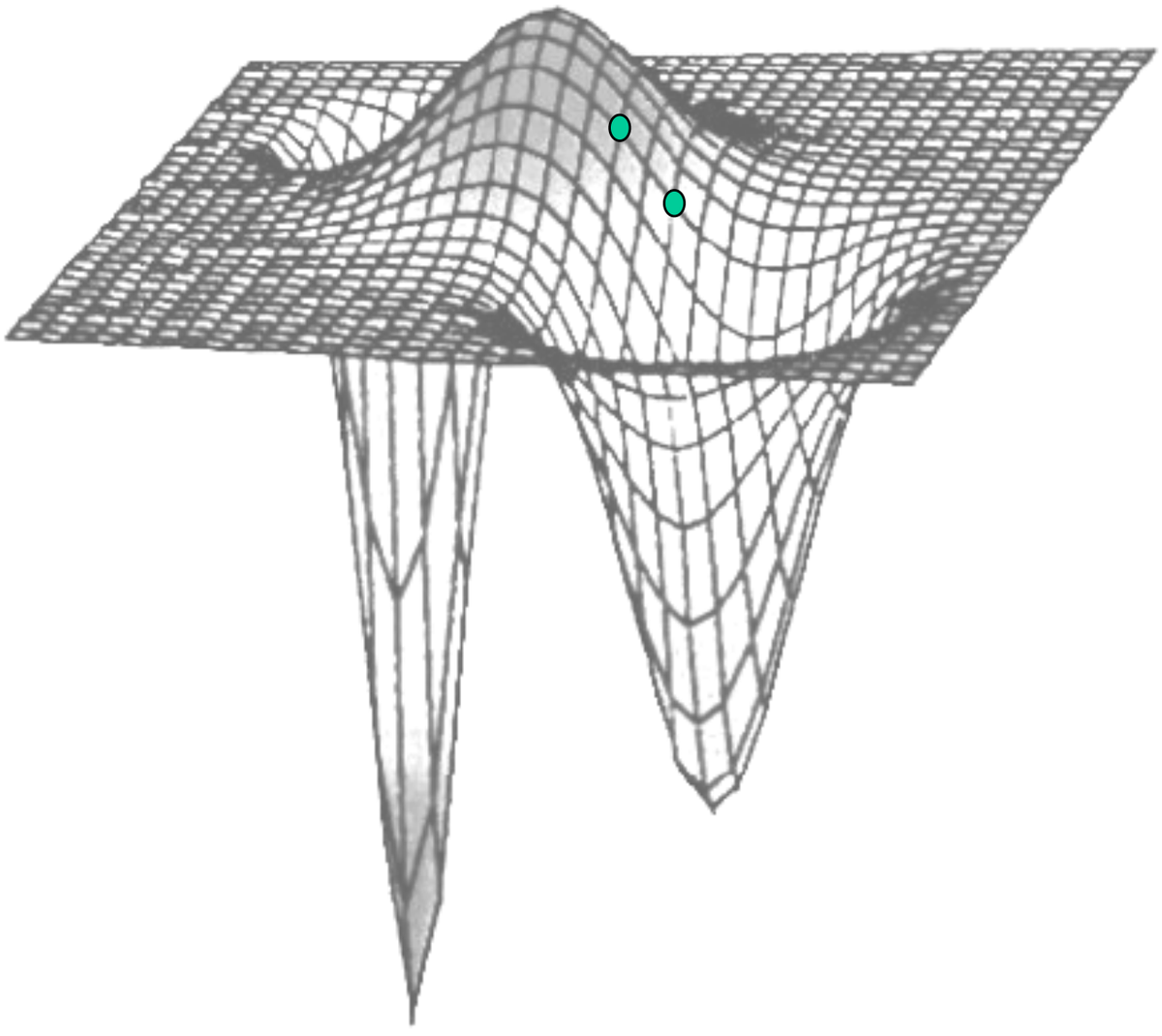


Neighbors in constitution space

- are in close chemical distance to each other
- are likely to be similar in their spectroscopic properties

Simulated Annealing

Guided Walk in Constitution Space

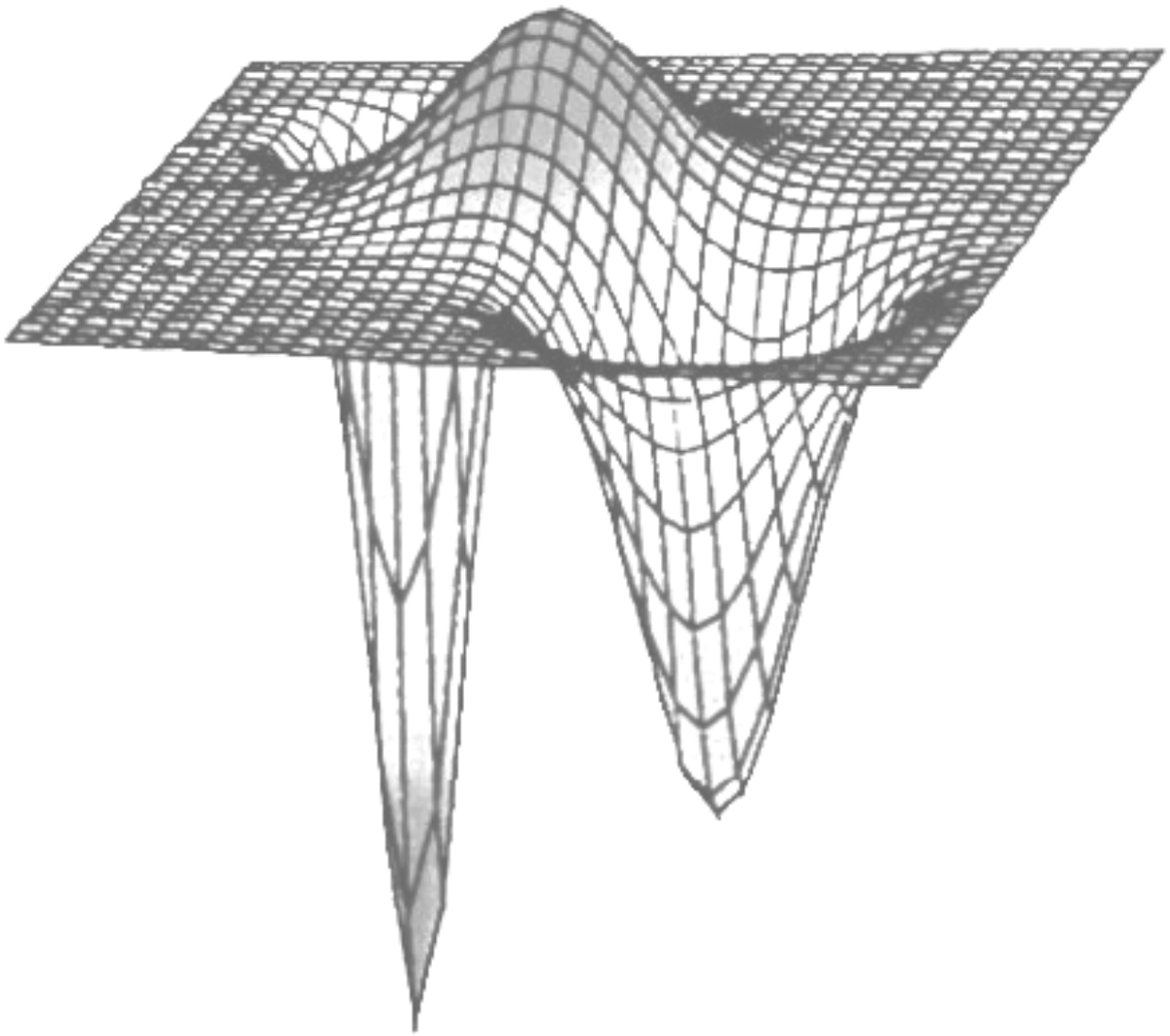


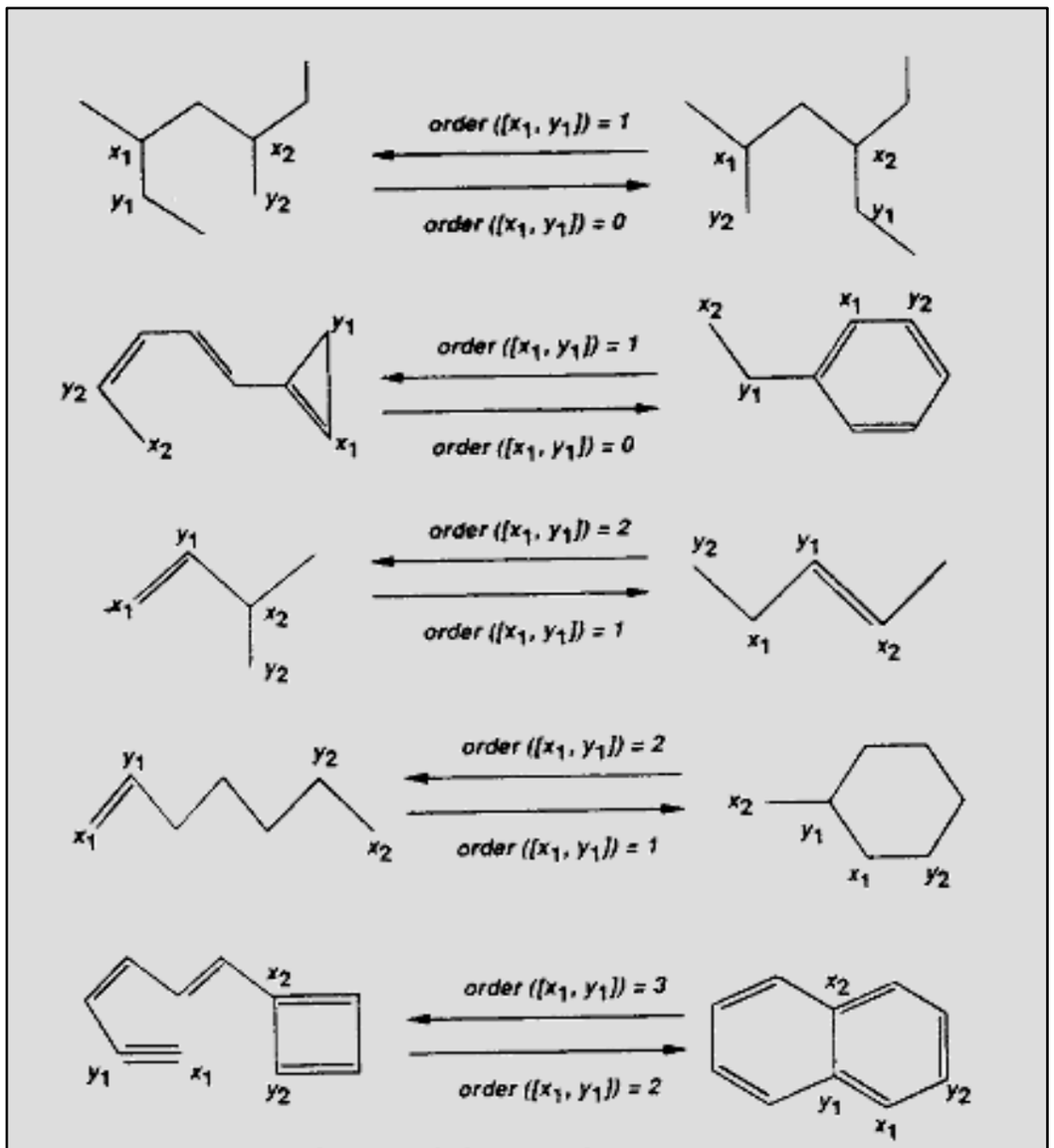
Neighbors in constitution space

- are in close chemical distance to each other
- are likely to be similar in their spectroscopic properties

Simulated Annealing

Small steps on the constitution space landscape

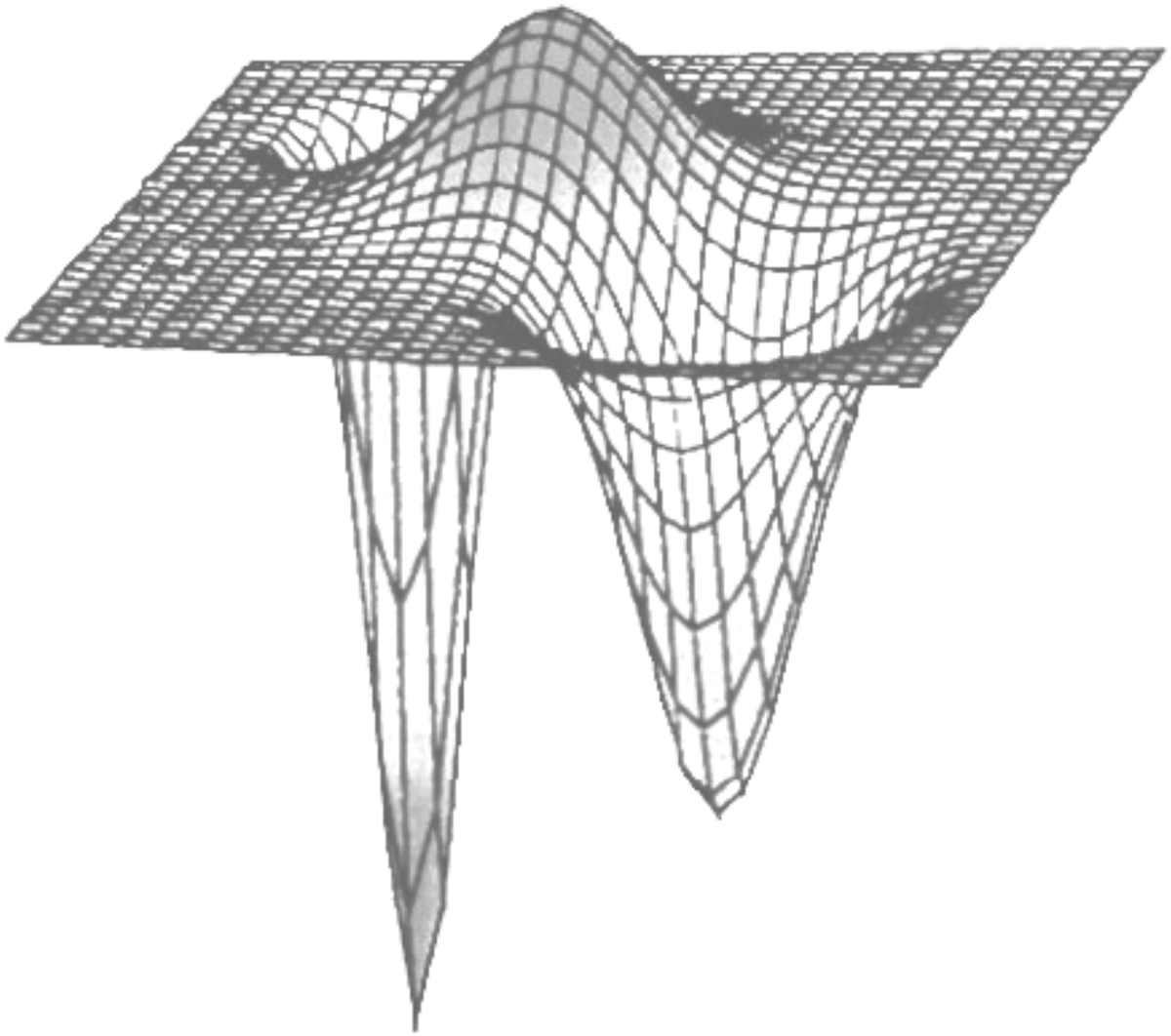




Faulon, J.-L.; J. Chem. Inf. Comput. Sci., 36 (1996) 4, 731-40

Simulated Annealing

Evaluating a score for each point (constitution)
in structure space

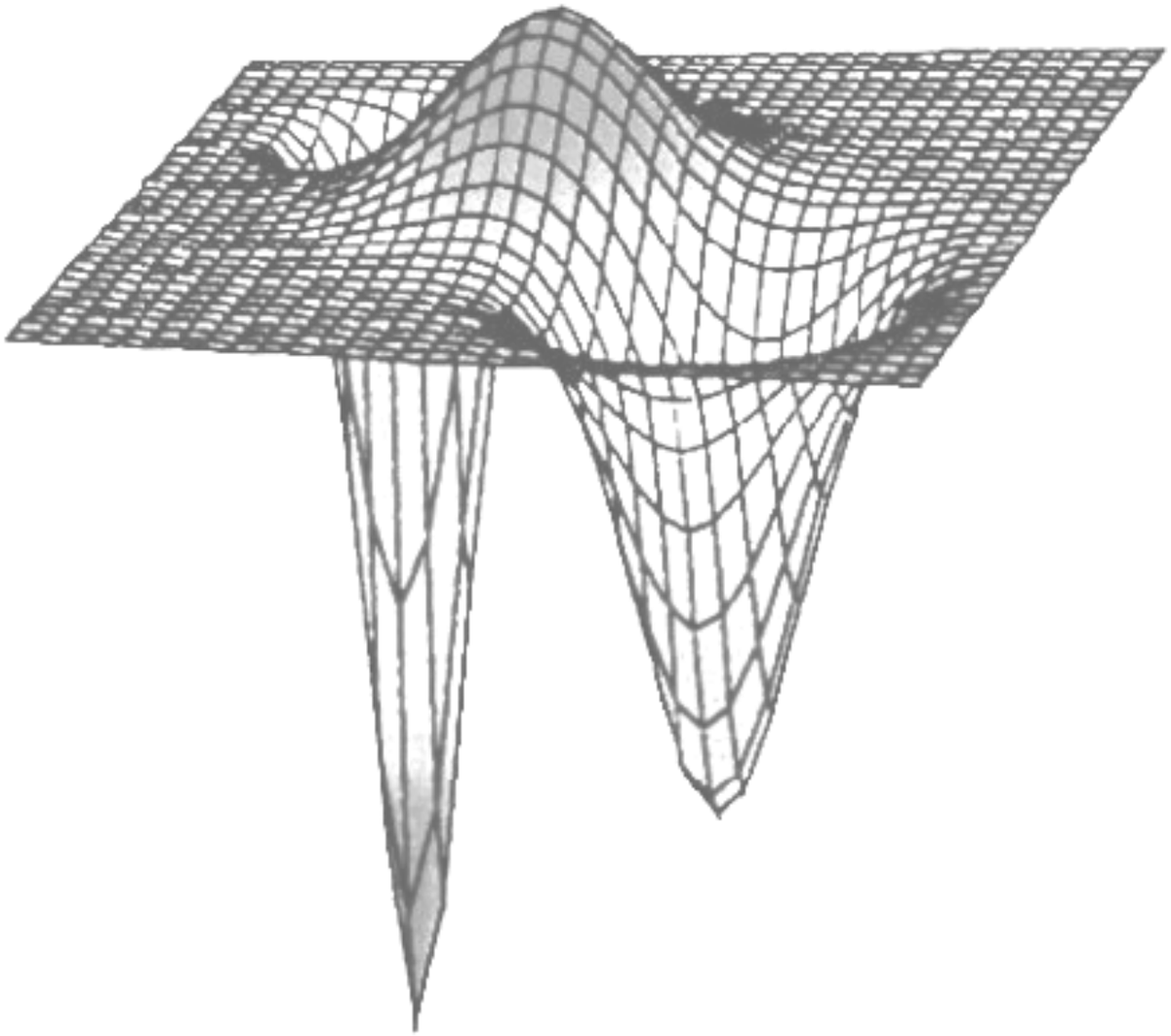


Score Function
based on
Spectroscopic Fitness

$$S_{total} =$$
$$c_1 S_{HMBC} +$$
$$c_2 S_{HHCOSY} +$$
$$c_3 S_{Shift} +$$
$$\ldots +$$
$$c_n S_{Features}$$

Simulated Annealing

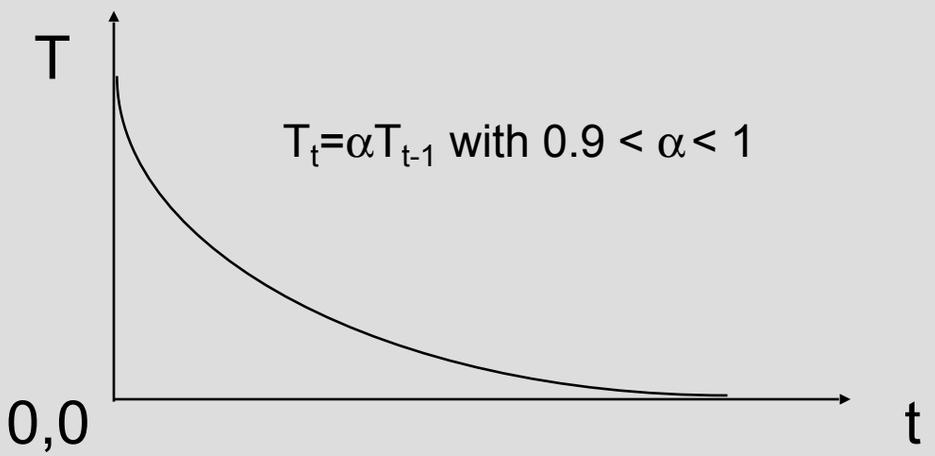
Small steps on the constitution space landscape



Acceptance criterion

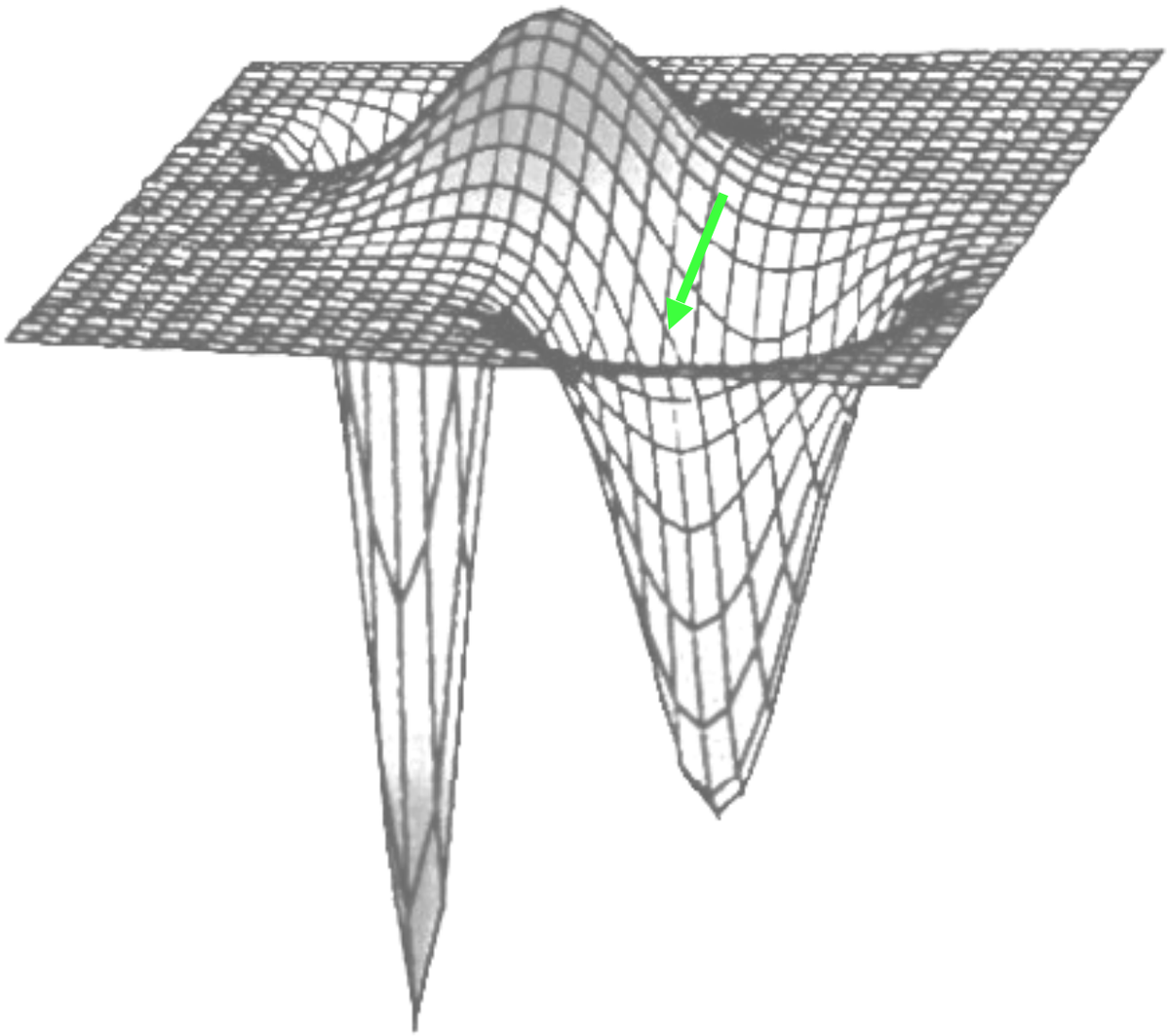
$$p = \exp(-\frac{\delta f}{T})$$

Annealing Schedule



Simulated Annealing

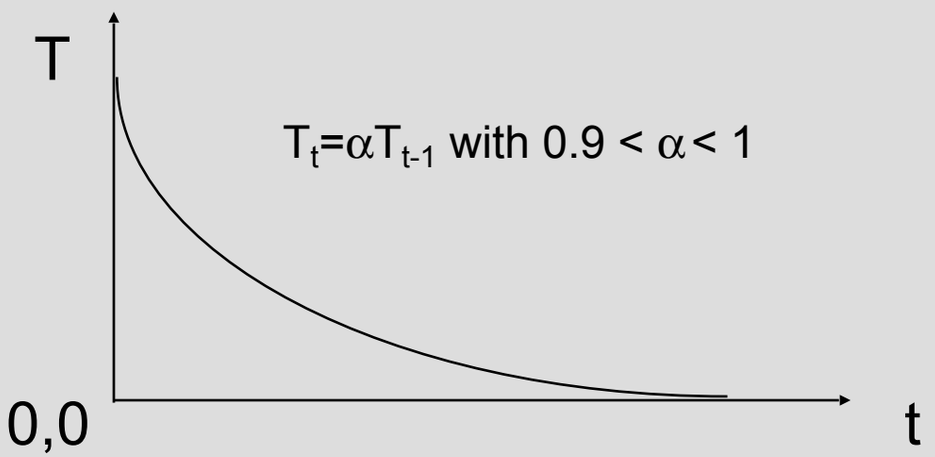
Small steps on the constitution space landscape



Acceptance criterion

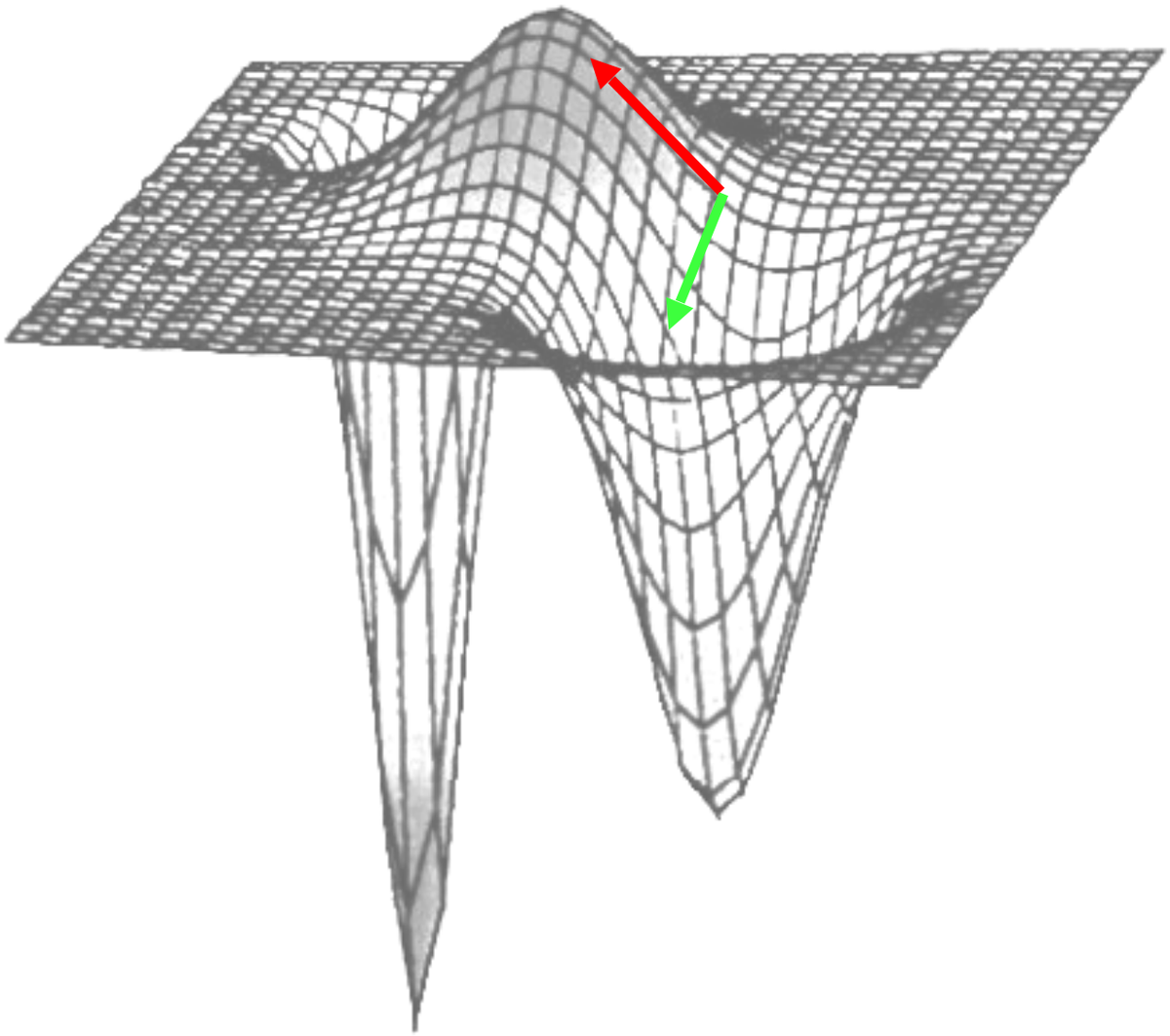
$$p = \exp(-\frac{\delta f}{T})$$

Annealing Schedule



Simulated Annealing

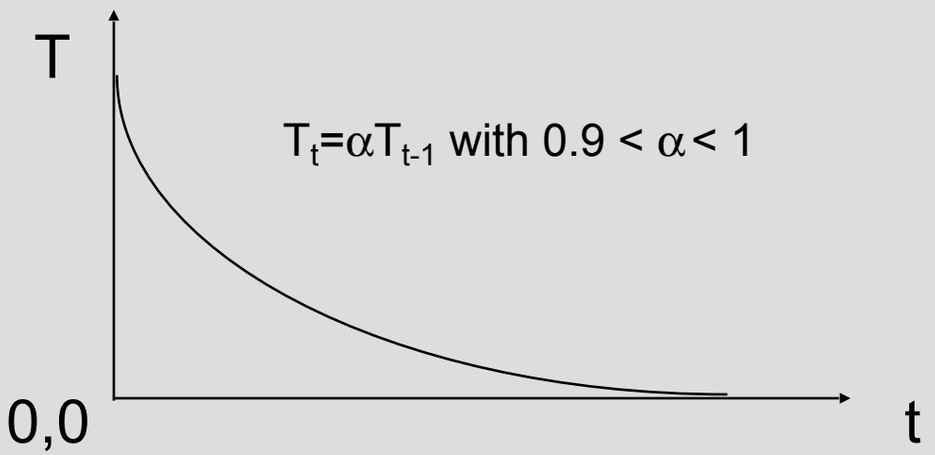
Small steps on the constitution space landscape



Acceptance criterion

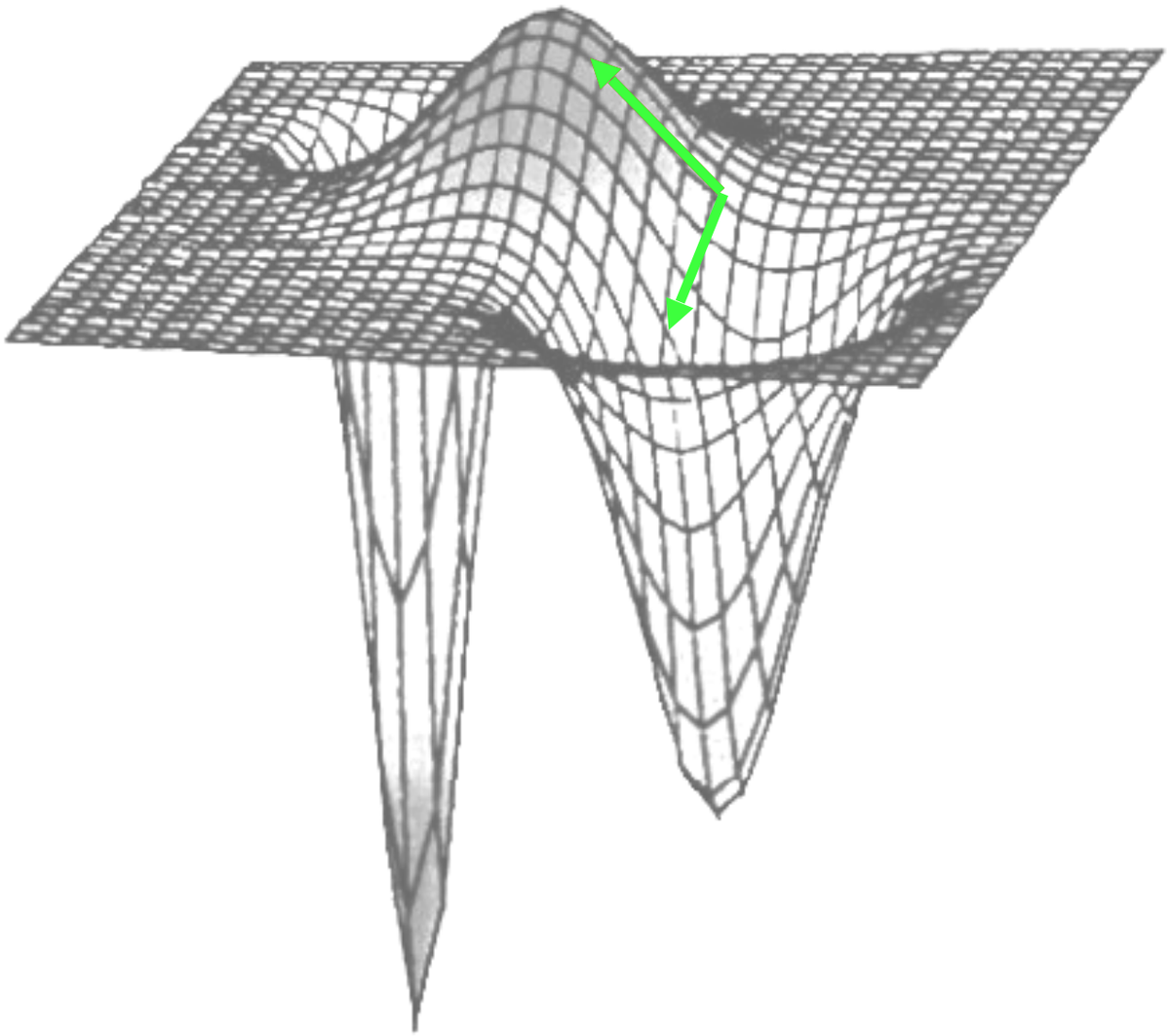
$$p = \exp(-\frac{\delta f}{T})$$

Annealing Schedule



Simulated Annealing

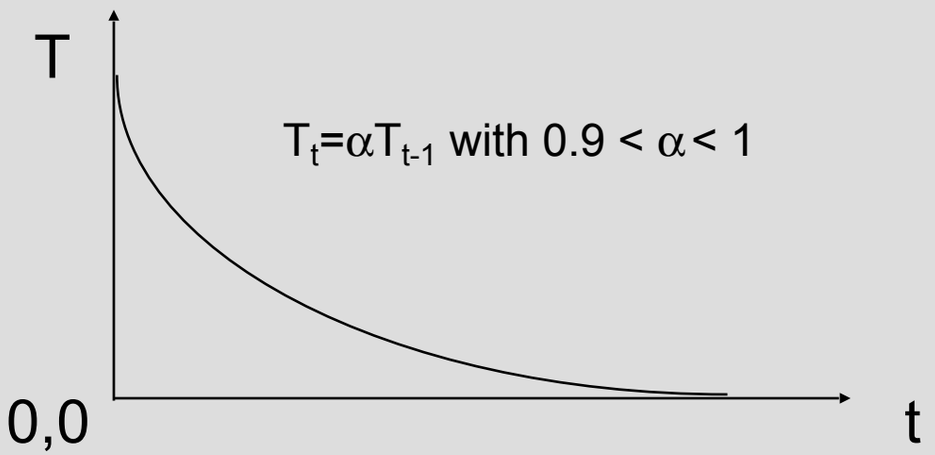
Small steps on the constitution space landscape



Acceptance criterion

$$p = \exp(-\frac{\delta f}{T})$$

Annealing Schedule



Some Nice Properties of this SA Scheme

Some Nice Properties of this SA Scheme

- Pluggable Target Function

Some Nice Properties of this SA Scheme

- Pluggable Target Function
  - If you can reliably calculate a measurable property for a given constitution, it can be part of your target function

Some Nice Properties of this SA Scheme

- Pluggable Target Function
  - If you can reliably calculate a measurable property for a given constitution, it can be part of your target function
    - Spectroscopic information
      - IR
      - UV-VIS
      - Other types of NMR experiments
      - MS fragmentation (?)

Some Nice Properties of this SA Scheme

- Pluggable Target Function
  - If you can reliably calculate a measurable property for a given constitution, it can be part of your target function
    - Spectroscopic information
      - IR
      - UV-VIS
      - Other types of NMR experiments
      - MS fragmentation (?)
    - Additional knowledge
      - Good-List/Bad-List fragments
      - Drug Likeness
      - Natural Product Likeness

Some Nice Properties of this SA Scheme

- Pluggable Target Function
  - If you can reliably calculate a measurable property for a given constitution, it can be part of your target function
    - Spectroscopic information
      - IR
      - UV-VIS
      - Other types of NMR experiments
      - MS fragmentation (?)
    - Additional knowledge
      - Good-List/Bad-List fragments
      - Drug Likeness
      - Natural Product Likeness

General System for Optimization in Constitution Space

Some Nice Properties of this SA Scheme

- Pluggable Target Function
  - If you can reliably calculate a measurable property for a given constitution, it can be part of your target function
    - Spectroscopic information
      - IR
      - UV-VIS
      - Other types of NMR experiments
      - MS fragmentation (?)
    - Additional knowledge
      - Good-List/Bad-List fragments
      - Drug Likeness
      - Natural Product Likeness

General System for Optimization in Constitution Space

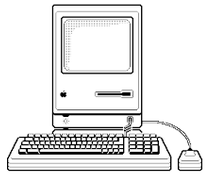
- Artifacts only lead to slightly lower ranking of correct structure in hit list

Distributed computing at its cheapest.

Gatekeeper

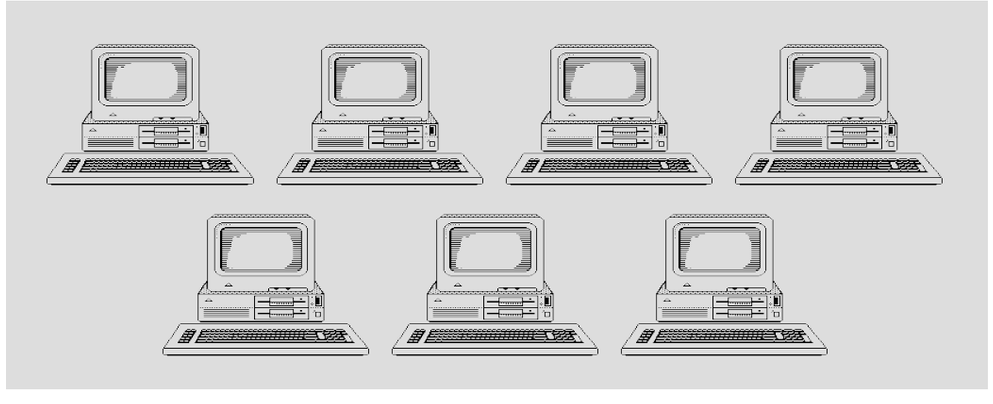
Retrieve Server List

Client

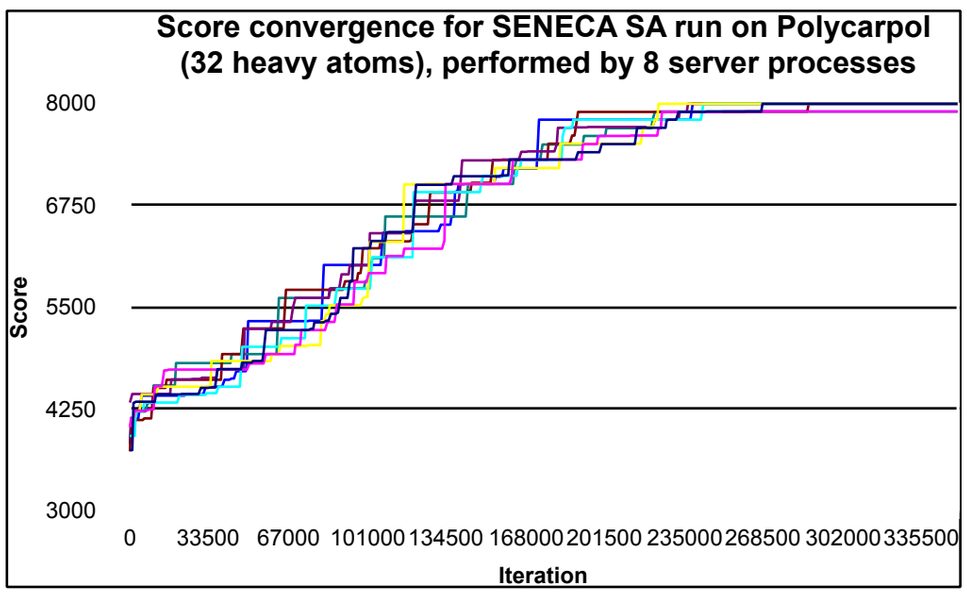


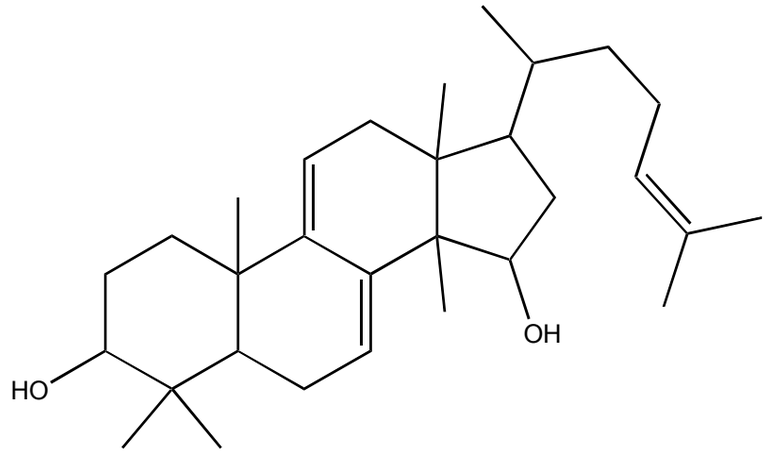
Submit Spectral Data

Collect Results



Distributed Server



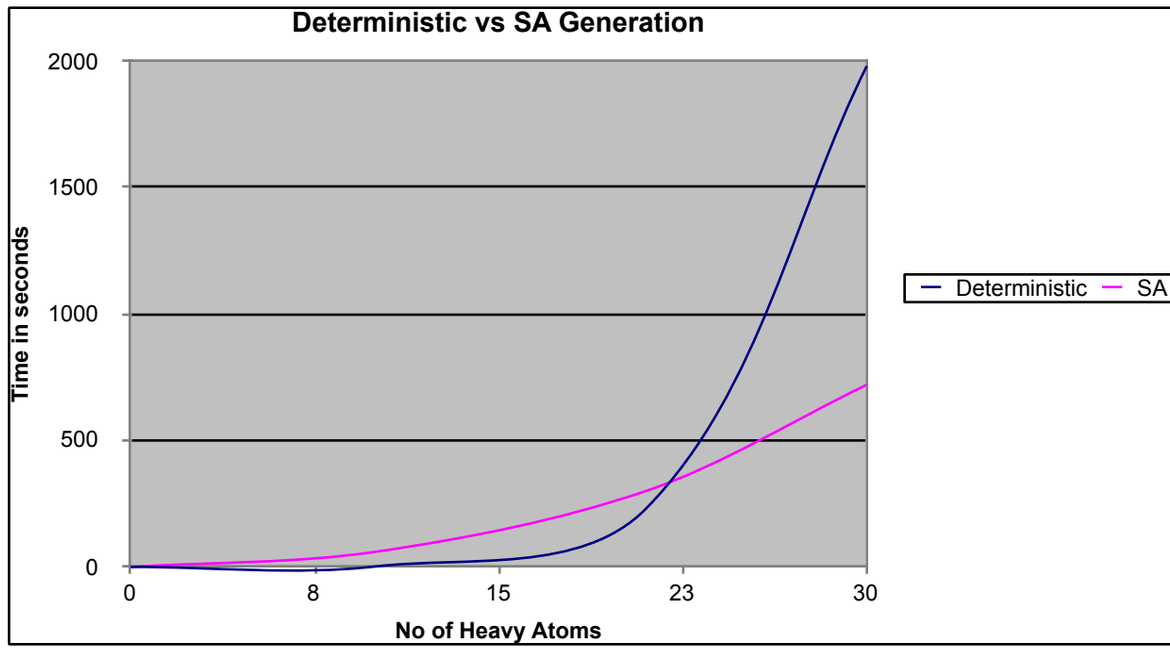


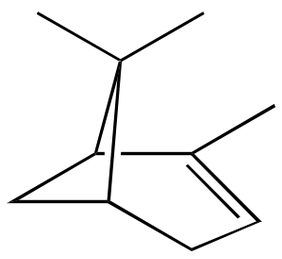
Polycarpol ($C_{30}H_{48}O_2$).

Qualitative assessment:

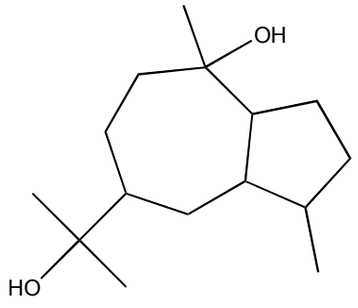
Computational complexity of deterministic and stochastic algorithms

Compound	LUCY	SENECA	Steps overall
α-Pinene ($C_{10}H_{16}$)	2 s	1 min	30,000
Eurabidiol($C_{15}H_{28}O_2$)	29 s	5 min	90 000
Polycarpol ($C_{30}H_{50}O$)	33 min	12 min	350,000

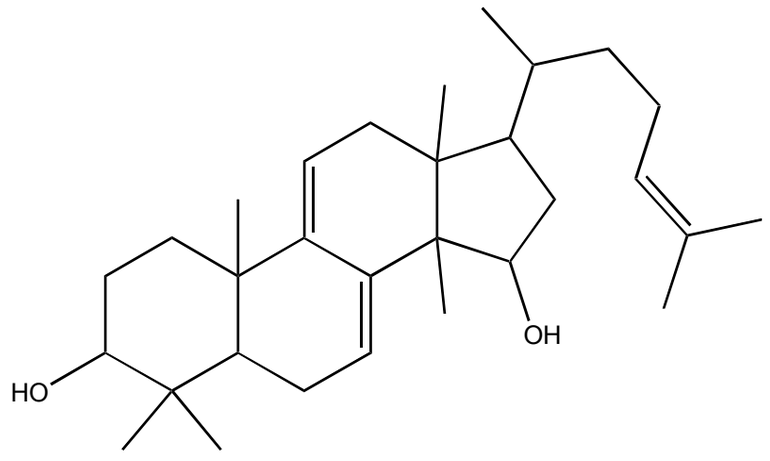




α-Pinene ($C_{10}H_{16}$)



Eurabidiol ($C_{15}H_{28}O_2$)



Polycarpol ($C_{30}H_{48}O_2$).

C. Steinbeck, *Journal of Chemical Information & Computer Sciences* **2001**, *41*, 1500.

Ranking Solutions

- Heteroatom-rich/proton-poor skeletons can yield many solutions (hundreds, thousands)
- Possible ranking by
  - Spectrum Similarity
  - Natural Product Likeness

Natural Product-likeness Score and Its Application for Prioritization of Compound Libraries

Peter Ertl,* Silvio Roggo, and Ansgar Schuffenhauer

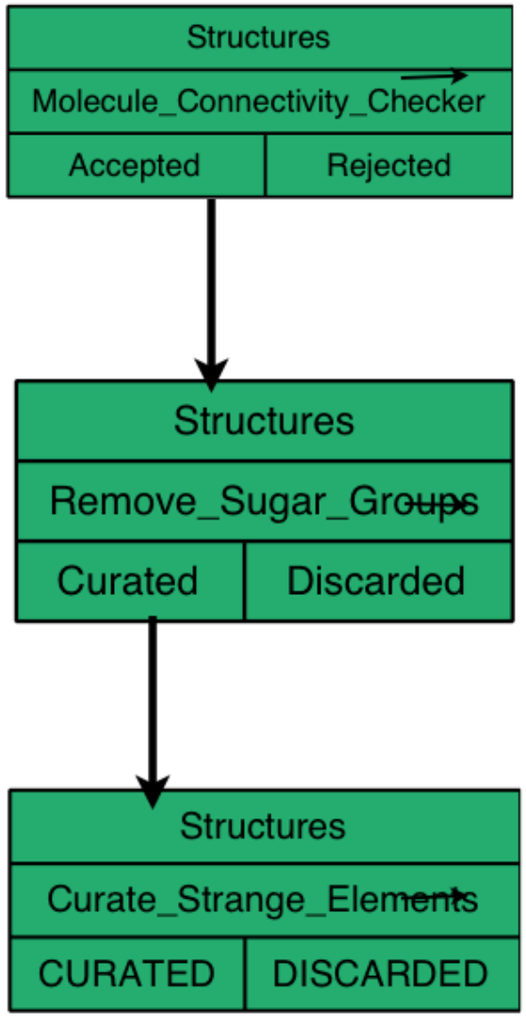
Novartis Institutes for BioMedical Research, CH-4002 Basel, Switzerland

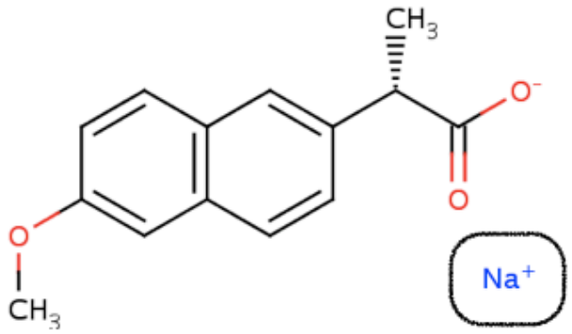
Datasets

Source	Number of molecules		
ChEMBL	22,876	TCM @TAIWAN	39162
ZINC	113,425		Training set
DrugBank	6634		
hmp	7531		
text mining	3182		Test set

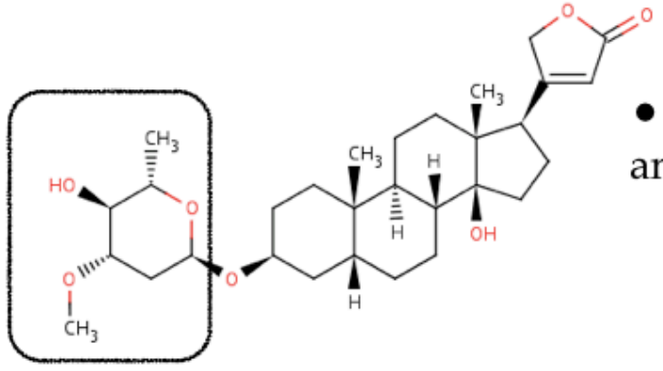
Components for Molecule Curation

Components for Molecule Curation

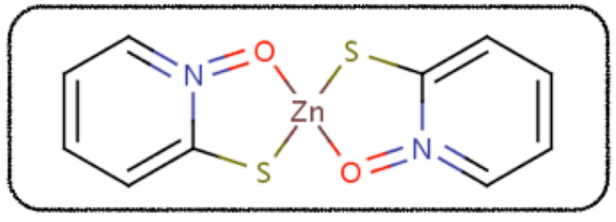




• This worker checks the connectedness of the molecule and removes counter ions and fragments less than 6 atoms.



• This worker removes from the scaffold linear sugars and ring sugars linked by glycosidic bond.

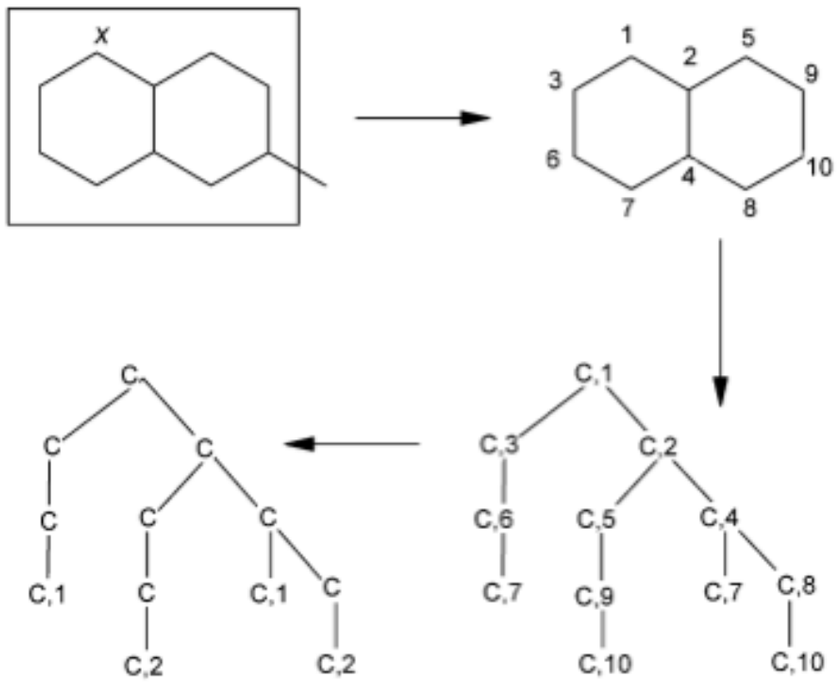


• This worker discards molecules completely if it contains elements other than non-metals (C, B, H, N, F, Cl, Br, I, O, S, P, Se and As).

The Signature Molecular Descriptor. 4. Canonizing Molecules Using Extended Valence Sequences

Jean-Loup Faulon,*,† Michael J. Collins,‡ and Robert D. Carr§

Departments of Computational Biology, Cryptography, and Discrete Algorithms and Mathematics, Sandia National Laboratories, P.O. Box 969, MS 9951, Livermore, California 94551, and Department of Computer Science MSC01 11301 1, University of New Mexico, Albuquerque, New Mexico 87131-0001



$^{0}\sigma(x)$ = C

$^{1}\sigma(x)$ = C (C C)

$^{2}\sigma(x)$ = C (C (C) C(C C))

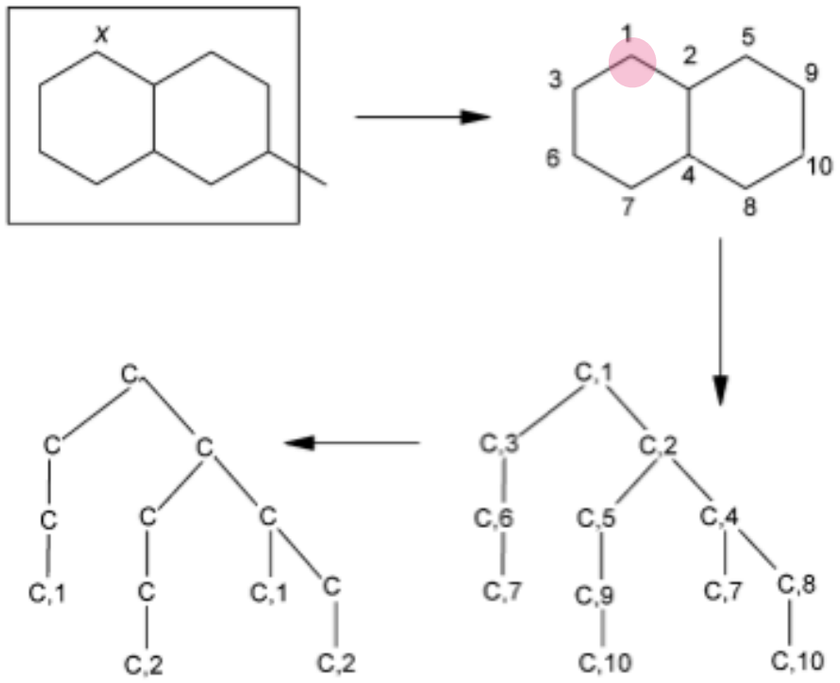
$^{3}\sigma(x)$ = C (C (C (C,1)) C(C (C) C (C,1 C)))

$^{4}\sigma(x)$ = C (C (C (C,1)) C(C (C (C,2)) C (C,1 C(C,2))))

The Signature Molecular Descriptor. 4. Canonizing Molecules Using Extended Valence Sequences

Jean-Loup Faulon,*,† Michael J. Collins,‡ and Robert D. Carr§

Departments of Computational Biology, Cryptography, and Discrete Algorithms and Mathematics, Sandia National Laboratories, P.O. Box 969, MS 9951, Livermore, California 94551, and Department of Computer Science MSC01 11301 1, University of New Mexico, Albuquerque, New Mexico 87131-0001



$^{0}\sigma(x)$ = C

$^{1}\sigma(x)$ = C (C C)

$^{2}\sigma(x)$ = C (C (C) C(C C))

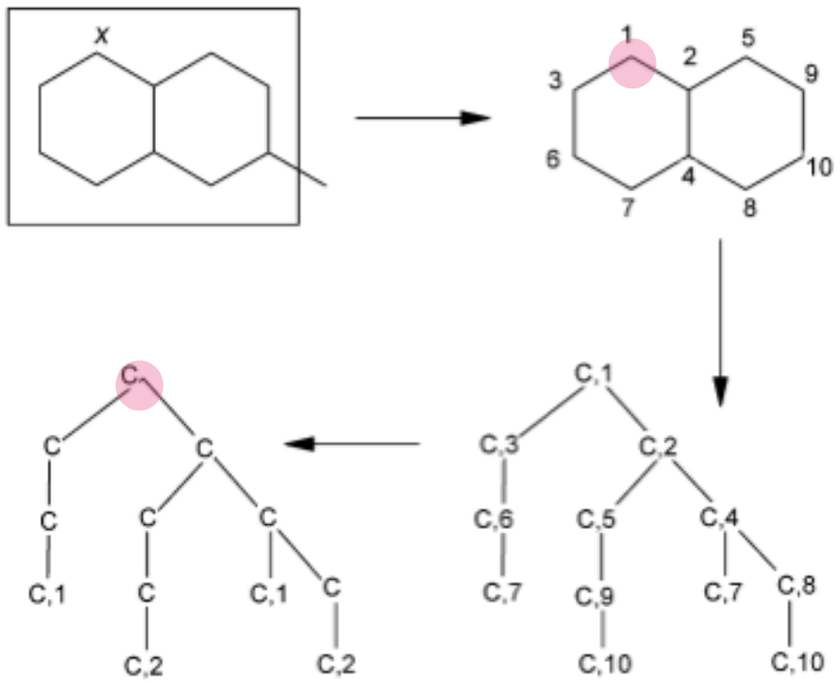
$^{3}\sigma(x)$ = C (C (C (C,1)) C(C (C) C (C,1 C)))

$^{4}\sigma(x)$ = C (C (C (C,1)) C(C (C (C,2)) C (C,1 C(C,2))))

The Signature Molecular Descriptor. 4. Canonizing Molecules Using Extended Valence Sequences

Jean-Loup Faulon,*,† Michael J. Collins,‡ and Robert D. Carr§

Departments of Computational Biology, Cryptography, and Discrete Algorithms and Mathematics, Sandia National Laboratories, P.O. Box 969, MS 9951, Livermore, California 94551, and Department of Computer Science MSC01 11301 1, University of New Mexico, Albuquerque, New Mexico 87131-0001



$^{0}\sigma(x)$ = C

$^{1}\sigma(x)$ = C (C C)

$^{2}\sigma(x)$ = C (C (C) C(C C))

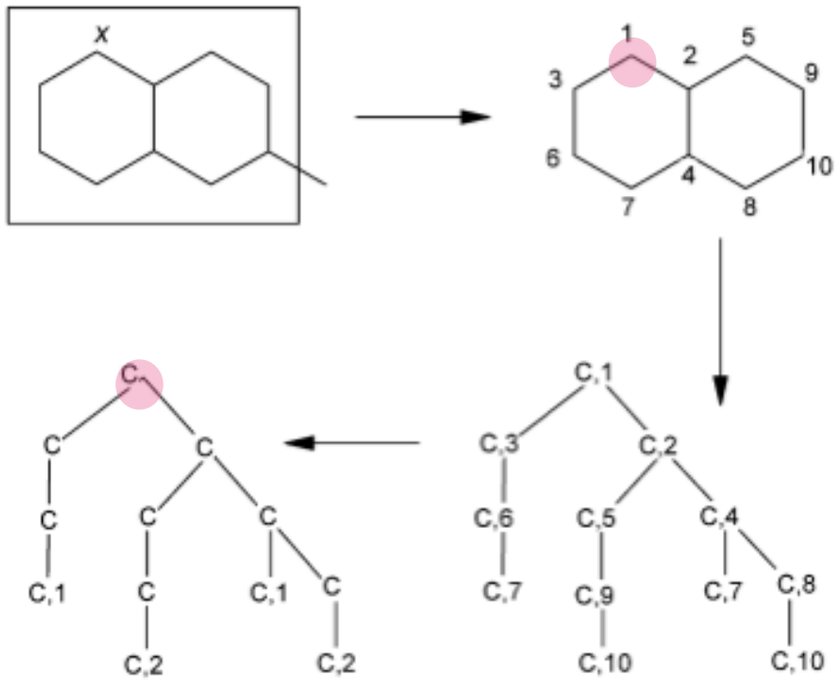
$^{3}\sigma(x)$ = C (C (C (C,1)) C(C (C) C (C,1 C)))

$^{4}\sigma(x)$ = C (C (C (C,1)) C(C (C (C,2)) C (C,1 C(C,2))))

The Signature Molecular Descriptor. 4. Canonizing Molecules Using Extended Valence Sequences

Jean-Loup Faulon,*,† Michael J. Collins,‡ and Robert D. Carr§

Departments of Computational Biology, Cryptography, and Discrete Algorithms and Mathematics, Sandia National Laboratories, P.O. Box 969, MS 9951, Livermore, California 94551, and Department of Computer Science MSC01 11301 1, University of New Mexico, Albuquerque, New Mexico 87131-0001



$^{0}\sigma(x)$ = C

$^{1}\sigma(x)$ = C (C C)

$^{2}\sigma(x)$ = C (C (C) C(C C))

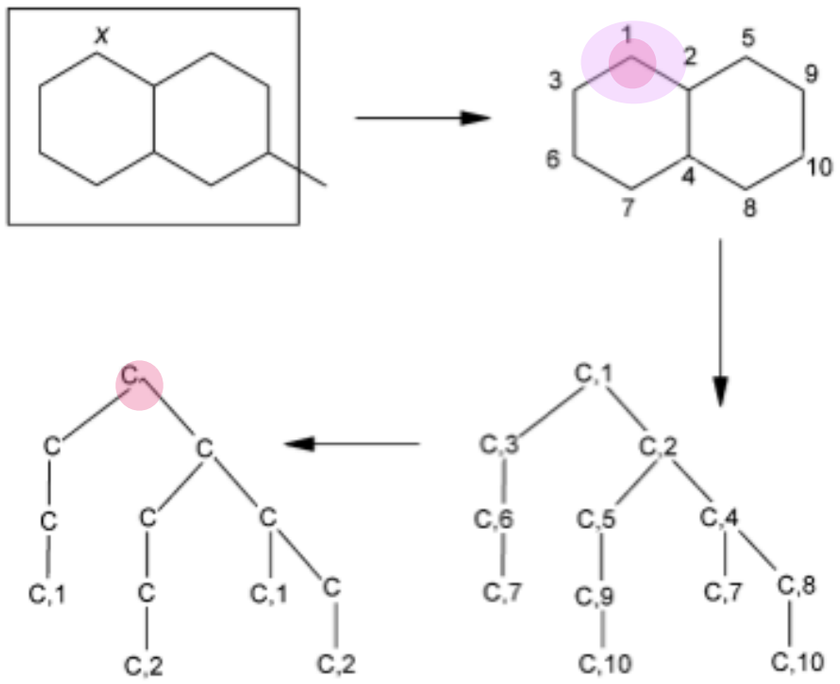
$^{3}\sigma(x)$ = C (C (C (C,1)) C(C (C) C (C,1 C)))

$^{4}\sigma(x)$ = C (C (C (C,1)) C(C (C (C,2)) C (C,1 C(C,2))))

The Signature Molecular Descriptor. 4. Canonizing Molecules Using Extended Valence Sequences

Jean-Loup Faulon,*,† Michael J. Collins,‡ and Robert D. Carr§

Departments of Computational Biology, Cryptography, and Discrete Algorithms and Mathematics, Sandia National Laboratories, P.O. Box 969, MS 9951, Livermore, California 94551, and Department of Computer Science MSC01 11301 1, University of New Mexico, Albuquerque, New Mexico 87131-0001



$^{0}\sigma(x)$ = C

$^{1}\sigma(x)$ = C (C C)

$^{2}\sigma(x)$ = C (C (C) C(C C))

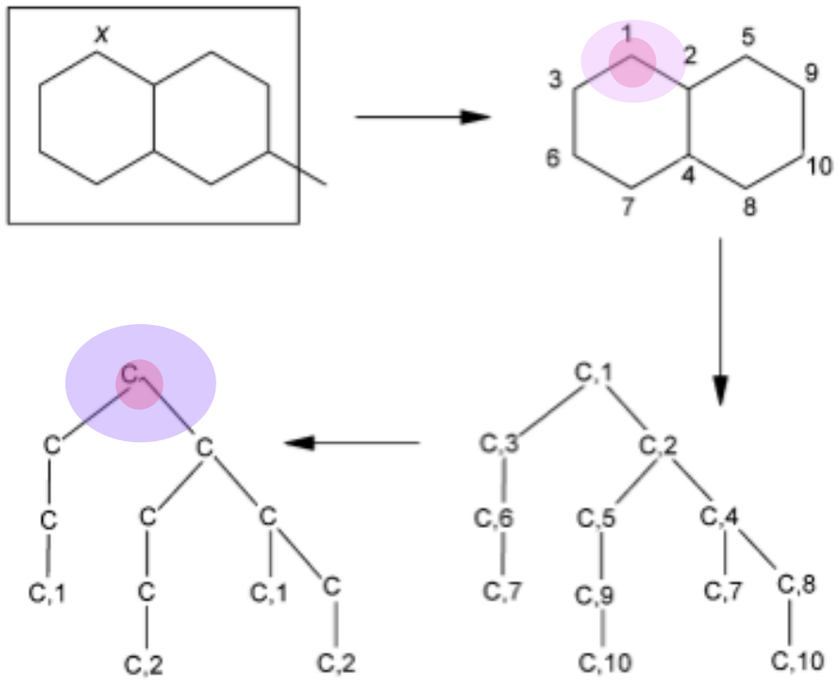
$^{3}\sigma(x)$ = C (C (C (C,1)) C(C (C) C (C,1 C)))

$^{4}\sigma(x)$ = C (C (C (C,1)) C(C (C (C,2)) C (C,1 C(C,2))))

The Signature Molecular Descriptor. 4. Canonizing Molecules Using Extended Valence Sequences

Jean-Loup Faulon,*,† Michael J. Collins,‡ and Robert D. Carr§

Departments of Computational Biology, Cryptography, and Discrete Algorithms and Mathematics, Sandia National Laboratories, P.O. Box 969, MS 9951, Livermore, California 94551, and Department of Computer Science MSC01 11301 1, University of New Mexico, Albuquerque, New Mexico 87131-0001



$^{0}\sigma(x)$ = C

$^{1}\sigma(x)$ = C (C C)

$^{2}\sigma(x)$ = C (C (C) C(C C))

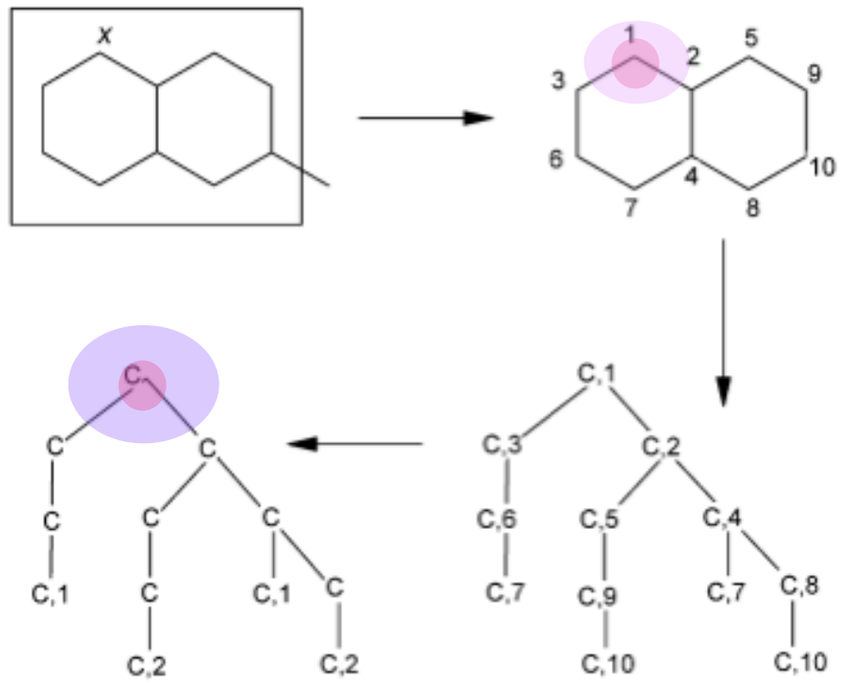
$^{3}\sigma(x)$ = C (C (C (C,1)) C(C (C) C (C,1 C)))

$^{4}\sigma(x)$ = C (C (C (C,1)) C(C (C (C,2)) C (C,1 C(C,2))))

The Signature Molecular Descriptor. 4. Canonizing Molecules Using Extended Valence Sequences

Jean-Loup Faulon,*,† Michael J. Collins,‡ and Robert D. Carr§

Departments of Computational Biology, Cryptography, and Discrete Algorithms and Mathematics, Sandia National Laboratories, P.O. Box 969, MS 9951, Livermore, California 94551, and Department of Computer Science MSC01 11301 1, University of New Mexico, Albuquerque, New Mexico 87131-0001



$^{0}\sigma(x)$ = C

$^{1}\sigma(x)$ = C (C C)

$^{2}\sigma(x)$ = C (C (C) C(C C))

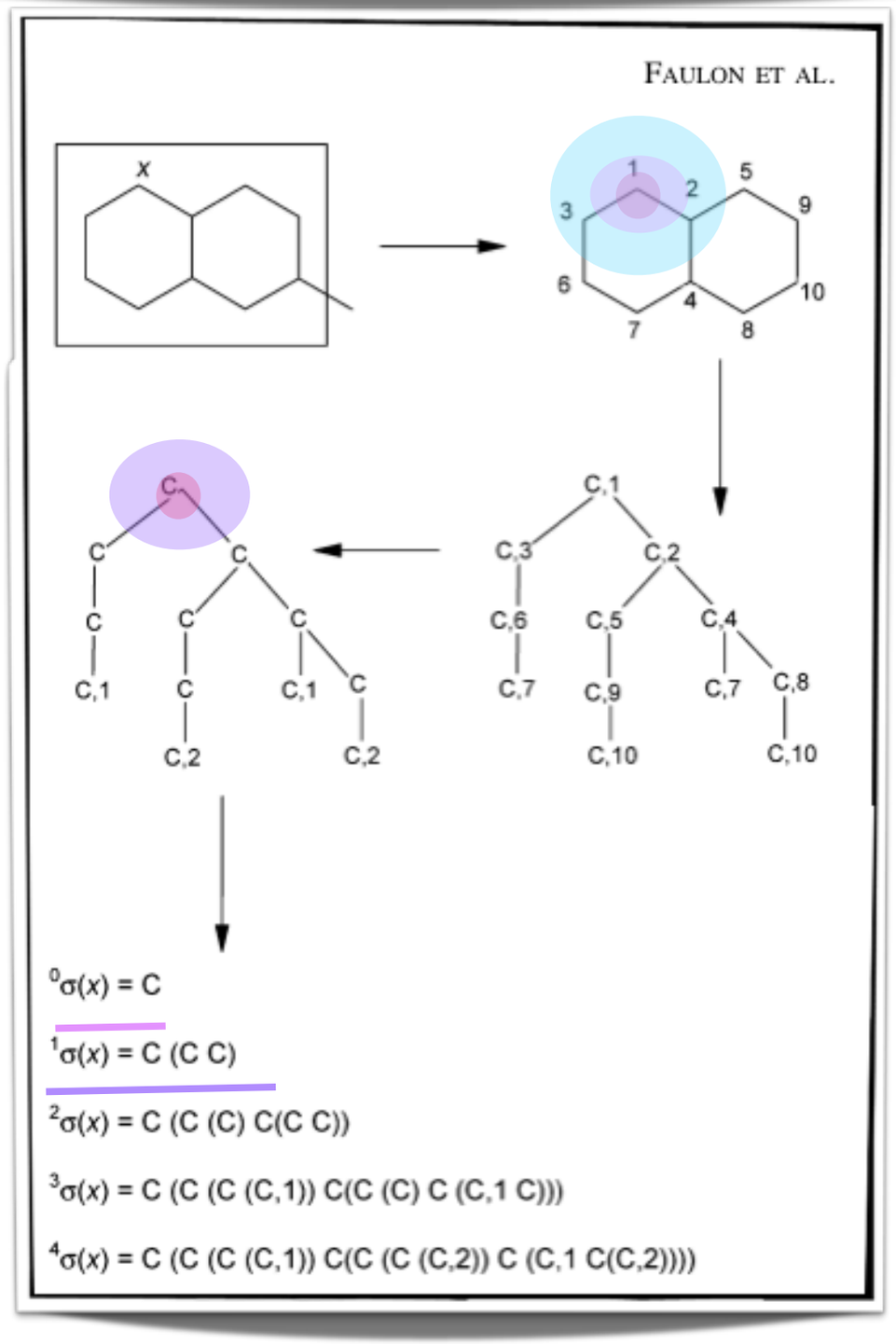
$^{3}\sigma(x)$ = C (C (C (C,1)) C(C (C) C (C,1 C)))

$^{4}\sigma(x)$ = C (C (C (C,1)) C(C (C (C,2)) C (C,1 C(C,2))))

The Signature Molecular Descriptor. 4. Canonizing Molecules Using Extended Valence Sequences

Jean-Loup Faulon,*,† Michael J. Collins,‡ and Robert D. Carr§

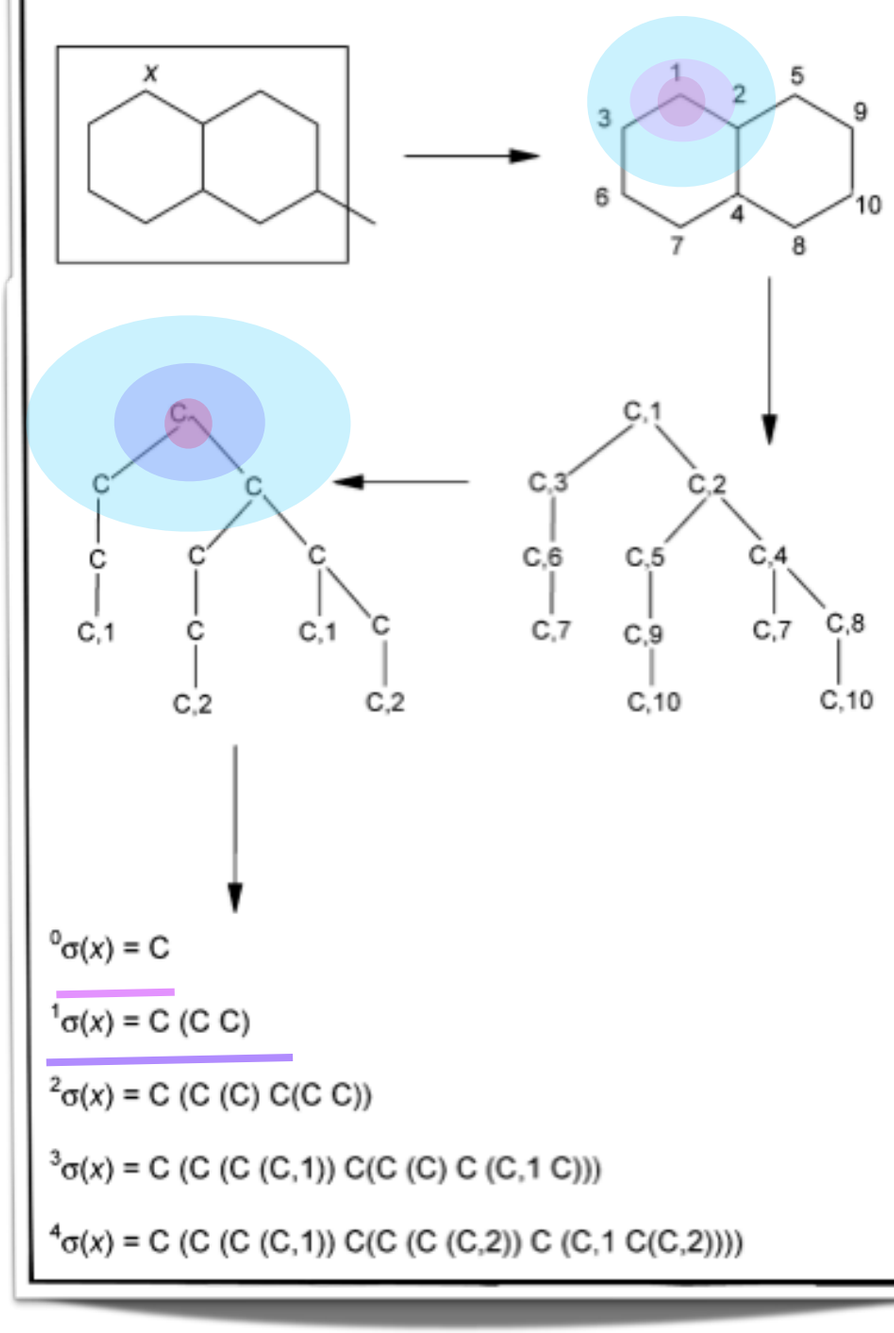
Departments of Computational Biology, Cryptography, and Discrete Algorithms and Mathematics, Sandia National Laboratories, P.O. Box 969, MS 9951, Livermore, California 94551, and Department of Computer Science MSC01 11301 1, University of New Mexico, Albuquerque, New Mexico 87131-0001



The Signature Molecular Descriptor. 4. Canonizing Molecules Using Extended Valence Sequences

Jean-Loup Faulon,*,† Michael J. Collins,‡ and Robert D. Carr§

Departments of Computational Biology, Cryptography, and Discrete Algorithms and Mathematics, Sandia National Laboratories, P.O. Box 969, MS 9951, Livermore, California 94551, and Department of Computer Science MSC01 11301 1, University of New Mexico, Albuquerque, New Mexico 87131-0001



$^{0}\sigma(x)$ = C

$^{1}\sigma(x)$ = C (C C)

$^{2}\sigma(x)$ = C (C (C) C(C C))

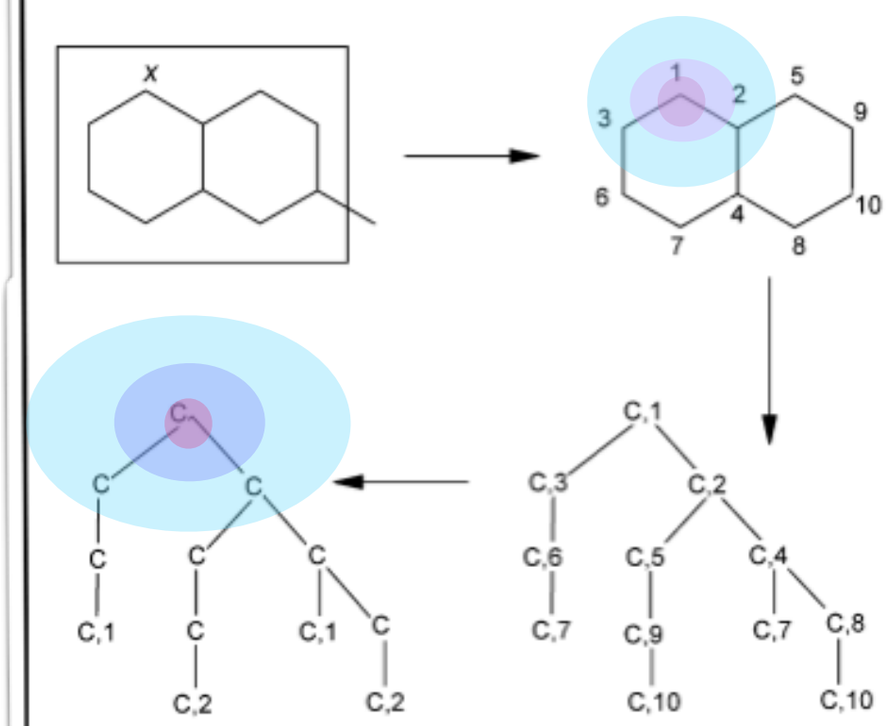
$^{3}\sigma(x)$ = C (C (C (C,1)) C(C (C) C (C,1 C)))

$^{4}\sigma(x)$ = C (C (C (C,1)) C(C (C (C,2)) C (C,1 C(C,2))))

The Signature Molecular Descriptor. 4. Canonizing Molecules Using Extended Valence Sequences

Jean-Loup Faulon,*,† Michael J. Collins,‡ and Robert D. Carr§

Departments of Computational Biology, Cryptography, and Discrete Algorithms and Mathematics, Sandia National Laboratories, P.O. Box 969, MS 9951, Livermore, California 94551, and Department of Computer Science MSC01 11301 1, University of New Mexico, Albuquerque, New Mexico 87131-0001



$^{0}\sigma(x)$ = C

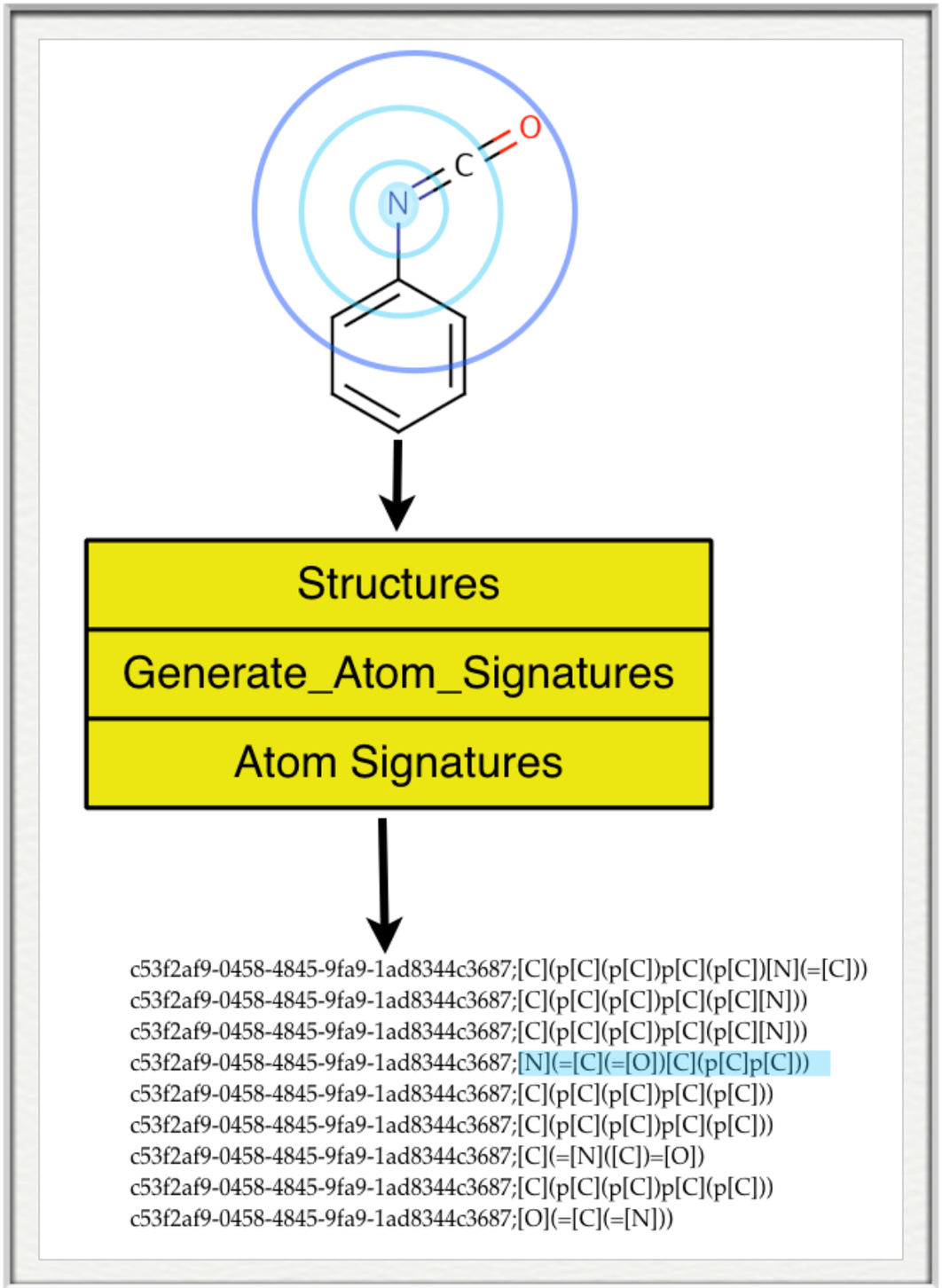
$^{1}\sigma(x)$ = C (C C)

$^{2}\sigma(x)$ = C (C (C) C(C C))

$^{3}\sigma(x)$ = C (C (C (C,1)) C(C (C) C (C,1 C)))

$^{4}\sigma(x)$ = C (C (C (C,1)) C(C (C (C,2)) C (C,1 C(C,2))))

Component for Signature Generation



Statistics for Scoring

Statistics for Scoring

$$Fragment_i = \log\left[\frac{NP_i}{SM_i} * \frac{SM_t}{NP_t}\right]$$

Statistics for Scoring

$$Fragment_i = \log\left[\frac{NP_i}{SM_i} * \frac{SM_t}{NP_t}\right]$$

NP - *Natural product*
SM - *Synthetic molecule*

Statistics for Scoring

$$Fragment_i = \log\left[\frac{NP_i}{SM_i} * \frac{SM_t}{NP_t}\right]$$

NP - *Natural product*
SM - *Synthetic molecule*

In the fragment contribution ($Fragment_i$),

✦NP_i is the total number of molecules in the NP dataset in which the $Fragment_i$ occurs,

✦SM_i is the total number of molecules in the SM dataset in which the $Fragment_i$ occurs,

✦SM_t is the total number of molecules int he SM dataset

✦NP_t is the total number of molecules in the NP dataset.

✦N is the number of fragments in given molecule

Statistics for Scoring

$$Fragment_i = \log\left[\frac{NP_i}{SM_i} * \frac{SM_t}{NP_t}\right]$$

NP - *Natural product*
SM - *Synthetic molecule*

$$Score_N = \sum_{i=0}^{N} Fragment_i$$

In the fragment contribution ($Fragment_i$),

✦NP_i is the total number of molecules in the NP dataset in which the $Fragment_i$ occurs,

✦SM_i is the total number of molecules in the SM dataset in which the $Fragment_i$ occurs,

✦SM_t is the total number of molecules int he SM dataset

✦NP_t is the total number of molecules in the NP dataset.

✦N is the number of fragments in given molecule

Statistics for Scoring

$$Fragment_i = \log\left[\frac{NP_i}{SM_i} * \frac{SM_t}{NP_t}\right]$$

NP - *Natural product*
SM - *Synthetic molecule*

In the fragment contribution (*Fragment*$_i$),

✦NP_i is the total number of molecules in the NP dataset in which the *Fragment*$_i$ occurs,

✦SM_i is the total number of molecules in the SM dataset in which the *Fragment*$_i$ occurs,

✦SM_t is the total number of molecules int he SM dataset

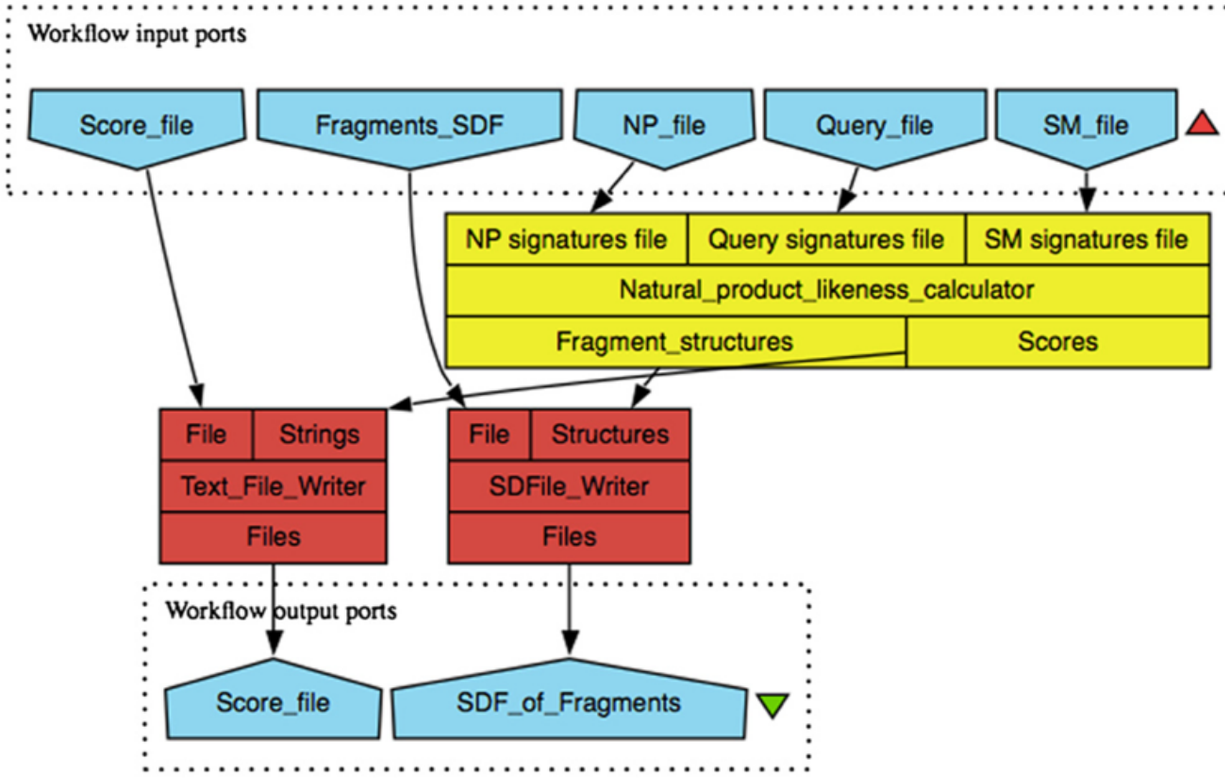
✦NP_t is the total number of molecules in the NP dataset.

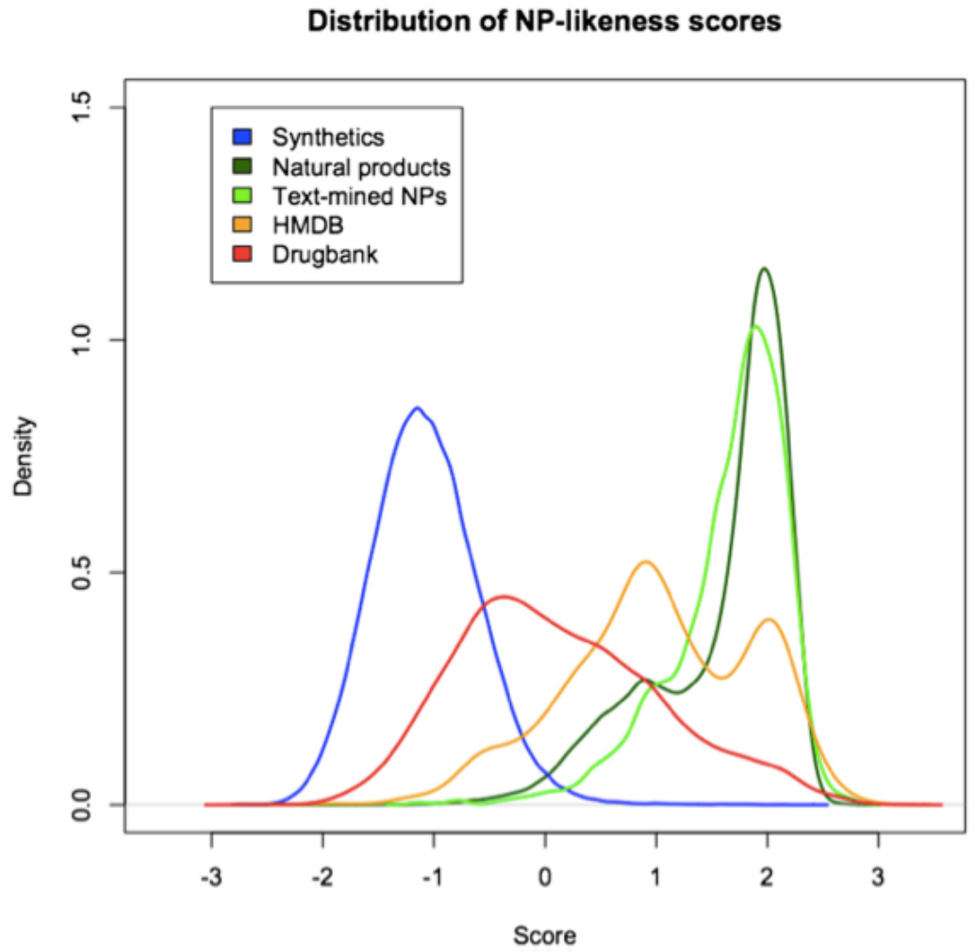
✦N is the number of fragments in given molecule

$$Score_N = \sum_{i=0}^{N} Fragment_i$$

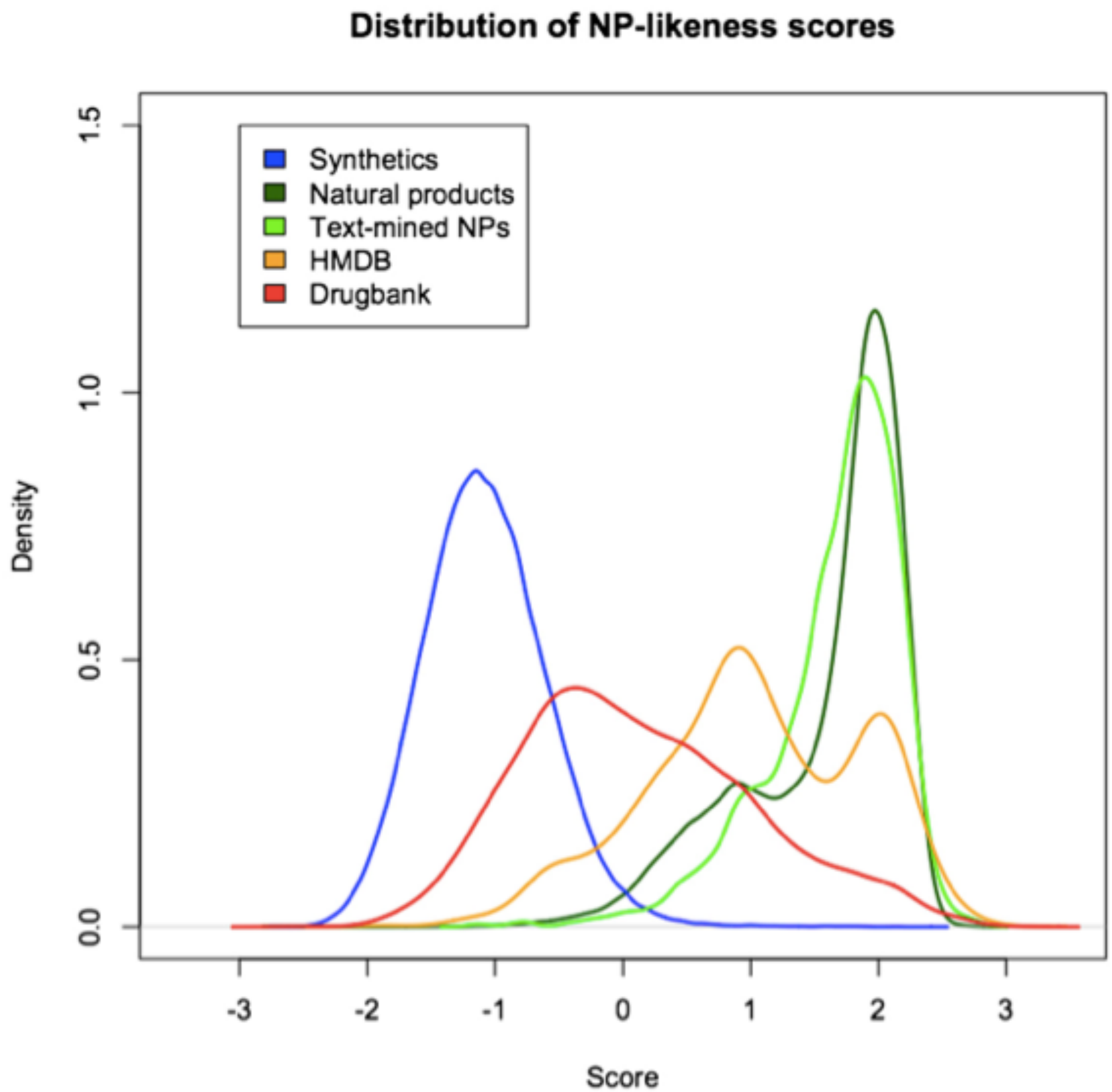
$$NormalisedScore = \frac{Score_N}{N}$$

- **Natural Product-likeness classification** and integrated it into Taverna workflow tool
  - (http://sourceforge.net/projects/np-likeness/).
- Included in second version of SENECA CASE





Jayaseelan KV, Moreno P, Truszkowski A, Ertl P & Steinbeck C (2012) Natural product-likeness score revisited: an open-source, open-data implementation. BMC Bioinformatics 13, 106.


Distribution of NP-likeness scores
Synthetics
Natural products
Text-mined NPs
HMDB
Drugbank
Density
0.0
0.5
1.0
1.5
-3
-2
-1
0
1
2
3
Score

Availability

Creator:

Kalai

Natural product likeness filter

Created: 13/05/11 @ 19:35:39 | **Last updated:** 29/08/11 @ 14:25:47

Reference: Natural Product-likeness Score and Its Application for Prioritization of Compound Libraries Peter Ertl,* Silvio Roggo, and Ansgar Schuffenhauer Novartis Institutes for BioMedical Research, CH-4002 Basel, Switzerland http://peter-ertl.com/reprints/Ertl-JCIM-48-68-2008.pdf
The natural product likeness scorer implemented by Peter Ertl was originally devised to screen large compound libraries for natural product likeness in drug designing studies. His work...

3 items in this pack

View

Download

myexperiment

Natural product likeness calculator Updated 2011-10-04

Calculates Natural Product(NP)-likeness of a molecule, i.e. the similarity of the molecule to the structure space covered by known natural products. NP-likeness is a useful criterion to screen compound libraries and to design new lead compounds.

sf Download
Natural Product Likeness c

sourceforge

Publicationen

Publicationen

New developments on the cheminformatics open workflow environment CDK-Taverna

Andreas Truszkowski[1], Kalai Vanii Jayaseelan[2], Stefan Neumann[3], Egon L Willighagen[4], Achim Zielesny[1]and Christoph Steinbeck[*2]

Publicationen

New developments on the cheminformatics open workflow environment CDK-Taverna

Andreas Truszkowski[1], Kalai Vanii Jayaseelan[2], Stefan Neumann[3], Egon L Willighagen[4], Achim Zielesny[1]and Christoph Steinbeck[*2]

Natural product-likeness score revisited: an open-source, open-data implementation

Kalai Vanii Jayaseelan[*1], Christoph Steinbeck[*1] , Pablo Moreno[1], Andreas Truszkowski[2], Peter Ertl[3]

The End