ELSEVIER PROPERTIES SA

Data: the Record of What we Think we Know, and Why

Sandy Lawson
Director R&D,
Elsevier Properties SA
Neuchâtel, Switzerland

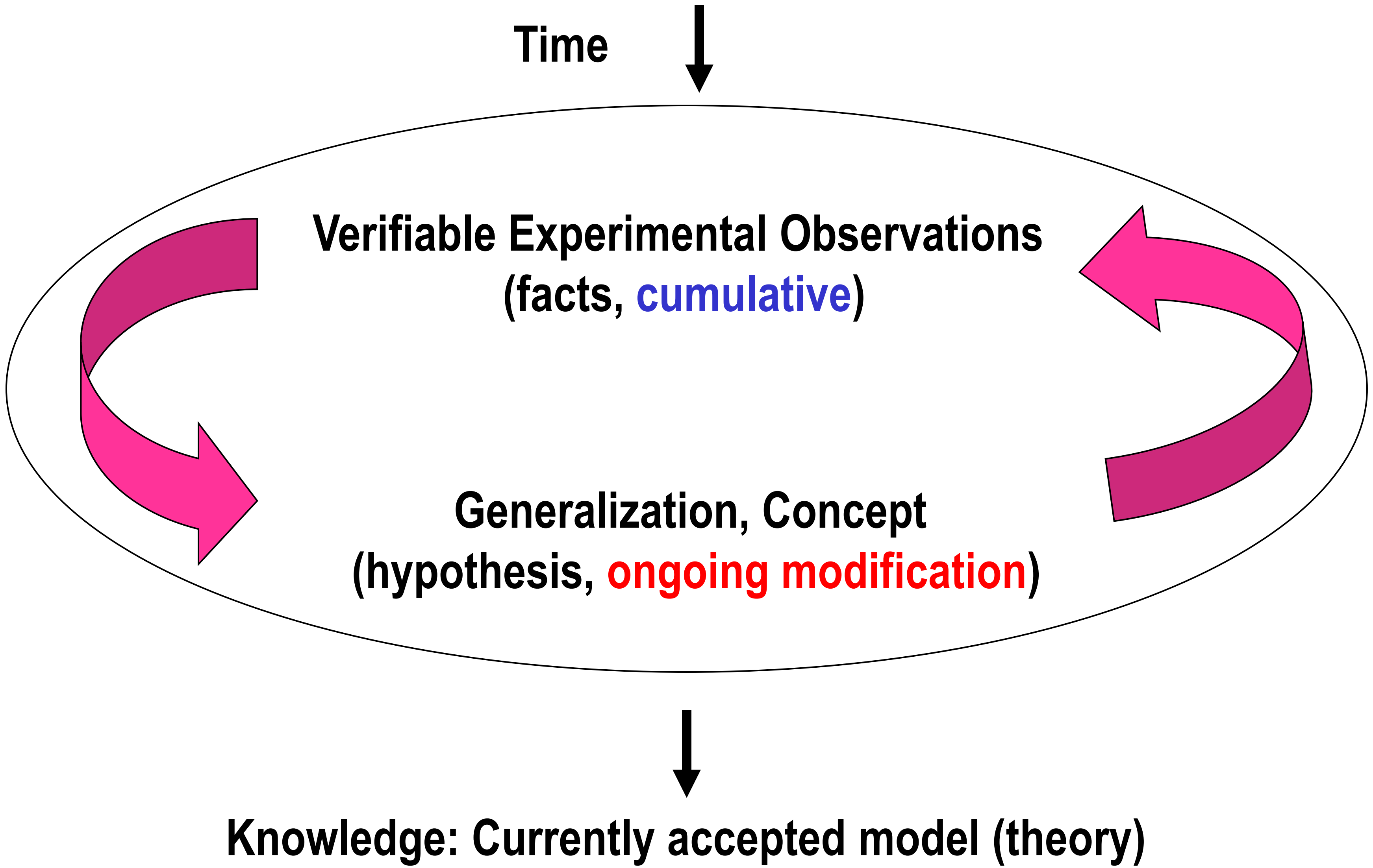
Caveat

- I am not a historian, I am an organic chemist with many years association with databases that primarily focus on organic chemistry.
- My views on the huge subject of “data” are to be seen in that light.
- Much of the historical background in this presentation was found on the internet (predominantly from Wikipedia, for which I acknowledge here with many thanks).

Data, Information and Knowledge are not clearly bounded

- Information (Facts)
  - Observed Data correlated under controlled conditions
    - » Based on what we think we **know**
- Hypotheses (Concepts)
  - What we **think** we know
    - » Based on the balance of reproducible information
- Knowledge (Theories)
  - What we think we **know**
    - » Standard Models, based on hypotheses not contradicted by reproducible information

The Scientific Method



All Data has its Hour in the Limelight

- 18th Century: Transition from Alchemy; Gravimetric data was key.
- 19th Century:
  - Physical Chemistry: Mathematical logic (e.g. Thermodynamics, Rudolf Clausius, 1850)
  - Organic Chemistry: Understanding structural principles
    - » Empirical logic: Building from something we **think** we know
      - Functional Group model

- 1900-1980:
  - Exploring Structure/Property Space (mechanisms, rules)
  - Identification methodologies (data used in analysis)
    - » Until 1960: colour, combustion analysis, melting point, refractive index, optical rotation, derivatives

- 1980-Present:
  - Useful Property data (Intrinsic and Extrinsic)
    - » Thermodynamics, boiling point, density, viscosity, solubility, pKs, azeotropes etc.
  - Ecological/Biological Data
    - » Environmental toxicology, physiological effects (e.g. IC50)
  - Reaction refinement
    - » Catalytic activity, regio- and stereo-selectivity (e.g. enantiomeric excess ratios)

Aristotle: 4th Century BC

- Animal, Vegetable or Mineral kingdoms
- Four elements: Earth, Air, Water, and Fire
  - ➔ Combination of two qualities:
    - » heat, cold, moisture, and dryness

  - Model dominated view of matter for nearly two thousand years.

- Robert Boyle 1661: “The Skeptical Chymist”
  - ➔ Met Galileo in 1641 as 14 year old. Asserted that his experiments denied the limiting of chemical elements to only the classic four; observed that metals gained weight when they burned. Formulated theory of atoms.

- Johann Joachim Becher 1667
  - ➔ Eliminated “Fire” and Air”, replaced by three forms of “Earth”

- Georg Stahl 1715
  - ➔ Phlogiston theory, derived from element “Fatty Earth”.
  - ➔ “Vitalism”:, organisms have a vital principle distinct from biochemical reactions
    - » Models dominated view of matter for nearly 100 years

Phlogiston and Vitalism were not Cranky Theories

- The Phlogiston theory explained burning, oxidation, calcination (metal residue after combustion), and breathing in the following way:
  - Flames extinguish because air becomes saturated with phlogiston.
  - Charcoal leaves little residue upon burning because it is nearly pure phlogiston.
  - Mice die in airtight space because air saturates with phlogiston.
  - When heated with charcoal, metals are restored because phlogiston is transferred from charcoal to the product calx (oxide).
  - *It could be all be true if a phlogiston atom had an atomic weight of -16*
- Joseph Black (1750): Discovery of CO_2 : by gravimetric **data**
  - **$MgCO_{3(s)} + H_2SO_4 \rightarrow MgSO_4 + CO_{2(g)} + H_2O$**
  - Did not require Phlogiston. Same gas as produced in life processes: the relation between combustion and metabolism as fundamentally related chemical processes (undermines Vitalism, precedes Wöhler). Father of Calorimetry.
- Mikhail Lomonosov (1756):
  - ➔ Contradicted Boyle: gravimetric **data**: without access of air from outside, the mass remains the same
- Joseph Priestley (1774): Discovery of O_2 : **$2HgO \rightarrow 2Hg + O_{2(g)}$**
  - "dephlogisticated air", which accounted for about 20% of atmospheric air

Phlogiston died in1775, Vitalism only retreated in 1828

- Antoine Lavoisier (1775)
  - Repeated Priestley's work with gravimetric **data,** developing a balance that could weigh to 0.0005g.
  - Proposed the Law of Conservation of Mass
  - "Furthermore, I repeat, in attacking here Stahl's doctrine my object is not to substitute a rigorously demonstrated theory but solely a hypothesis which appears to me more probable, more conformable to the laws of nature, and which appears to me to contain fewer forced explanations and fewer contradictions."
  - ➔ Priestley died in 1804 still believing in the Phlogiston theory.

- Friedrich Wöhler (1828): synthesised urea from inorganic components
  - He subsequently wrote to Berzelius that he had witnessed "The great tragedy of science, the slaying of a beautiful hypothesis by an ugly fact".
  - ➔ Louis Pasteur (1858): Rejecting the claims of Berzelius, Liebig,Traube and others that fermentation resulted from chemical agents or catalysts within cells, he concluded that fermentation was a "vital action".
  - Vitalism has fallen out of favour, though it had advocates even into the twentieth century.
  - ➔ Hans Driesch (1867–1941), an eminent embryologist
  - ➔ Most shoppers would still feel that Vitamin C from oranges was intrinsically not the same as synthetic Vitamin C

Why did Phlogiston suffer a quicker Demise than Vitalism?

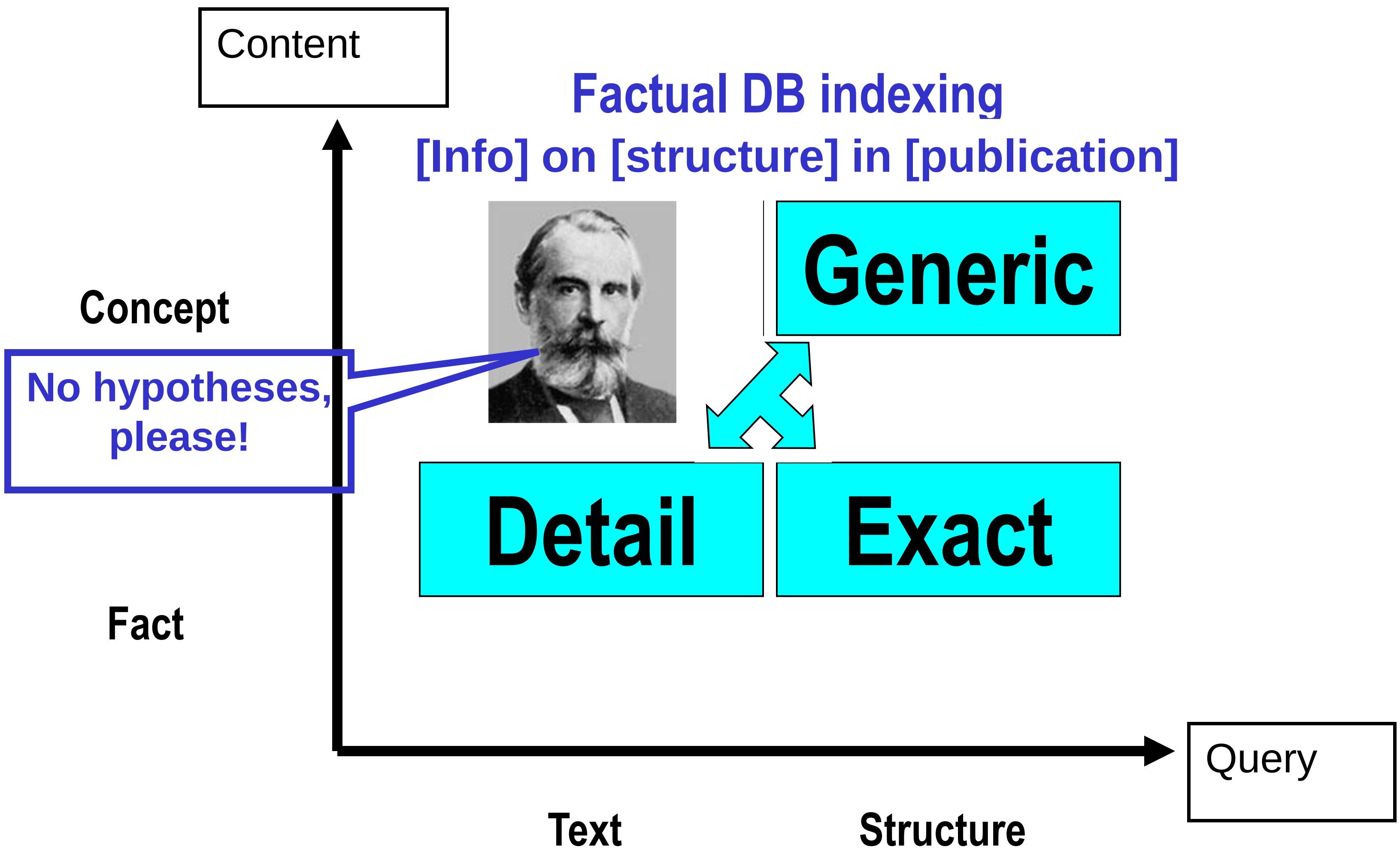
- It wasn't because vitalism was resistant to experiment *per se*: many vitalists were in fact accomplished experimentalists, including most notably Pasteur and Driesch. One aspect is that vitalism offers no definite predictions, and therefore cannot be proved wrong.
- But in my view, much more to the point is that gravimetric data is quantitative, i.e. expressed in numbers. Numerically-based hypotheses are simply more convincing, because the associated data are easier to measure, easier to refute.
- In a sense, with Mendeleev, the 19th century saw the first cementation (in chemistry) of the numerical basis that had brought the understanding of physics (and indeed physical chemistry) so far in the previous centuries.
- Yet chemistry remains to a large extent a science based on concepts that are difficult or (near-impossible) to express numerically. It is a science still with profound roots in *analogous thought*, rather than digital expression, and the concept of the Functional Group was cemented with the work of Beilstein.

- **That aspect had notable effect on the development of information services in the late 19th and early 20th centuries.**

Representative Major Database Providers

- **CAS:** Traditionally seen as a comprehensive bibliographic database of *concepts* in the general area of chemistry
  - Lately also collecting gene sequences, calculated numerical values and spectral shifts
- **CrossFire Beilstein:** Traditionally seen as a comprehensive *factual* database in organic chemistry
  - Lately also collecting pharmacological data, abstracts and other softer data

Factual and Bibliographic DBs



19th Century : Understanding Structural Principles

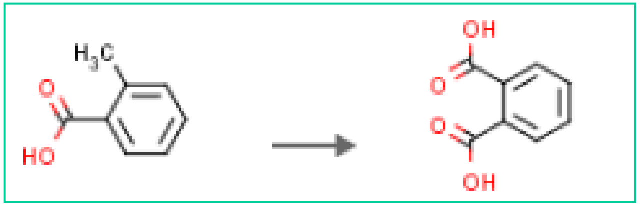
- 1881 Beilstein was criticized for adopting ideas of Kekulé (Liebermann)
- All critically evaluated property data listed in entry
  - Reactions were listed under “properties” of the product

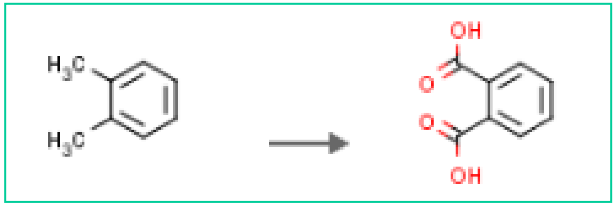
- Beilstein developed a unique, numerical structure sort based on Functional Groups, whose influence on nomenclature is underestimated today.
- Analogous structures were located in nearby entries

“Compounds in Beilstein are arranged according to the "Beilstein System," a complex series of rules which permits every substance to be assigned one specific location based on its structure”.
2010 :University of Texas at Austin
http://www.lib.utexas.edu/chem/info/beilstein.html

Compound Type	Volumes	System Numbers
Acyclic (no ring)	1-4	1-449
Isocyclic (all-carbon ring(s))	5-16	450-2358
Heterocyclic (other ring(s))	17-27	2359-4720

19th Century : Understanding Structural Principles





Claus; Pieszcek	1886	Chemische Berichte, **1886**, vol. 19, p. 3084

For over 100 years, synthesis from a known material was the ultimate method for proof of structure.

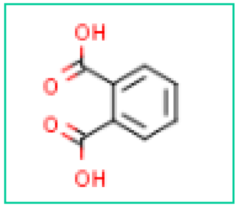
Then ... **R.B Woodward: "The Total Synthesis of Strychnine" (Tetrahedron 1963, 247-288):** (commenting on Robinson's great work over decades on the structure elucidation, compared with the short and easy period required for the same task with the new tool of X-Ray crystal structure analysis)

This short history should give pause to those whose talent for despair is lavished upon an organic chemistry ornamented and supplemented—or as they fancy, burdened—by magnificent new tools which permit the establishment in days or weeks of enlightenments which once would have required months or years. While it is undeniable that organic chemistry will be deprived of one special and highly satisfying kind of opportunity for the exercise of intellectual *élan* and experimental skill when the tradition of purely chemical structure elucidation declines, it is true too that the not infrequent dross of such investigation will also be shed; nor is there any reason to suppose that the challenge for the hand and the intellect must be less, or the fruits less tantalizing, when chemistry *begins* at the advanced vantage point of an established structure.

Of course, men make much use of excuses for activities which lead to discovery, and the lure of unknown structures has in the past yielded a huge dividend of unsought fact, which has been of major importance in building organic chemistry as a science. Should a surrogate now be needed, we do not hesitate to advocate the case for synthesis.

20th Century : Exploring Property Space

(Example taken from Reaxys, November 2010)



Pre 1960

Physical Data

- Melting Point (52)
- Density of the Liquid (2)
- Adsorption (MCS) (29)
- Association (MCS) (42)
- Boundary Surface Phenomena (MCS) (3)
- Conformation (1)
- Crystal Phase (6)
- Crystal Property Description (4)
- Crystal System (1)
- Decomposition (1)
- Density of the Crystal (3)
- Dissociation Exponent (313)
- Electrical Data (13)
- Electrical Moment (4)
- Electrochemical Behaviour (21)
- Electrochemical Characteristics (6)
- Enthalpy of Combustion (6)
- Enthalpy of Formation (4)
- Enthalpy of Fusion (1)
- Enthalpy of Sublimation (3)
- Enthalpy of Vaporization (1)
- Further Information (2)
- Heat Capacity Cp (2)
- Interatomic Distances and Angles (2)
- Ionization Potential (1)
- Liquid Phase (1)
- Liquid/Liquid Systems (MCS) (21)
- Liquid/Solid Systems (MCS) (9)
- Liquid/Vapour Systems (MCS) (4)
- Magnetic Susceptibility (4)
- Mechanical & Physical Properties (MCS) (2)
- Mechanical Properties (2)
- Molecular Deformation (1)
- Optics (3)
- Other Thermochemical Data (9)
- Partition octan-1-ol/water (MCS) (1)
- Solubility (MCS) (261)
- Solubility Product (MCS) (1)
- Solution Behaviour (MCS) (2)
- Space Group (2)
- Surface Tension (1)
- Triple Point (1)
- Vapour Pressure (1)

1960

Spectra

- NMR Spectroscopy (38)
- IR Spectroscopy (25)
- Mass Spectrometry (12)
- UV/VIS Spectroscopy (26)
- Raman Spectroscopy (11)
- Luminescence Spectroscopy (5)
- Fluorescence Spectroscopy (5)
- Phosphorescence Spectroscopy (2)
- Other Spectroscopic Methods (1)

1995

Bioactivity/Ecotox

- Pharmacological Data (53)
- Ecotoxicology (24)
- Exposure Assessment (6)
- Concentration in the Environment (39)
- Transport and Distribution (5)
- Biodegradation (12)
- Abiotic Degradation, Hydrolysis (1)
- Abiotic Degradation, Photolysis (9)
- Oxygen Demand (3)

Use/Application

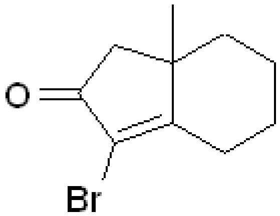
Natural Product

Copper(II) and Cadmium(II) Sorption onto Ferrihydrite in the Presence of Phthalic acid: Some Properties of the Ternary Complex	Song, Yantao; Swedlund, Peter J.; Singhal, Naresh	2008	Environmental Science and Technology, **2008**, vol. 42, # 11 p. 4008 - 4013 Full Text View citing articles

1980

- Title/Abstract
- Show All Substances (1)

Functional Group Models: 1941, Woodward's Rules (Empirical)

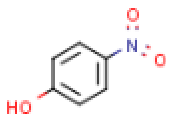


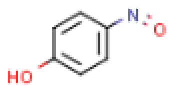
Five-membered ring parent enone:		202 nm
a-Br:		25 nm
b-Alkyl groups or ring residues:	2 x 12 =	24 nm
Exocyclic double bond:		5 nm
Calculated:		256 nm
Observed:		251 nm

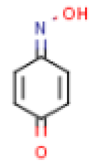
http://www.chemistry.ccsu.edu/glagovich/teaching/316/uvvis/conjugated.html

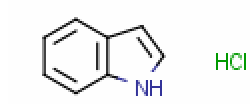
MO Models: 1965, Woodward-Hoffmann Rules; Orbital Symmetry

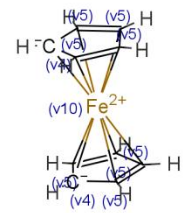
Most organic chemists still think in terms of valence bonds, functional groups and structural analogies based on them. This has some unfortunate consequences in the area of representation in chemical diagrams.











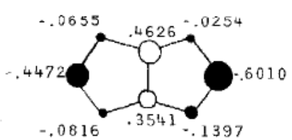
Frontier orbital controlled cycloaddition of 2-azapentalenes

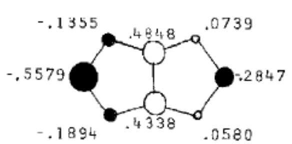
Klaus Hafner, *, Heinz-Gerd Kläs and Michael C. Böhm
Institut für Organische Chemie der Technischen Hochschule, Petersenstr. 22, D-6100 Darmstadt Germany
Received 12 October 1979. Available online 5 March 2001.

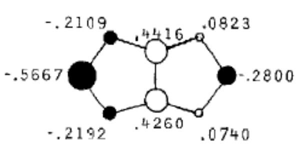
Abstract
It is shown that cycloadditions of 1,3-bis (dimethylamino) pentalenes and their 2-aza-analogues with activated alkines are frontier orbital controlled whereas protonation is charge controlled.

Article Outline

• References







Tetrahedron Letters
Volume 21, Issue 1, 1980, Pages 41-44

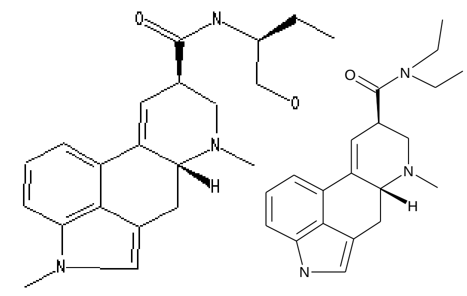
21st Century: Analogy in Structure/Property Relationships e.g. Methysergide

Lysergic Acid Diethyl Amide (LSD)

Field Availability List 1-2

Code	Field Name	Occ.
PHARM	Bioactivity: Pharmacological Data	57
CNR	Reference	41

Q02: Hit 5



Pharmacological Data 1 of 57

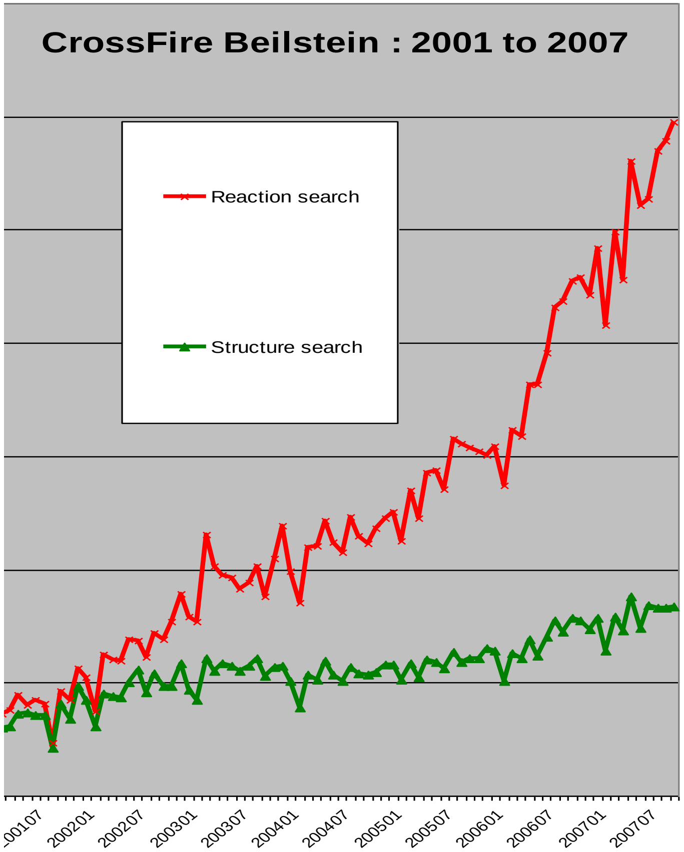
Effect	agonist
Species or Test-System	HEK293 cells
Method	stimulation of phosphatidylinositol hydrolysis via h5-HT$_{2B}$ receptors activation determined
Further Details	reference comp.: 5-HT (5-hydroxytryptamine), EC50: 1 nmol/l
Type	EC50
Value of Type	150 nmol/l
Ref. 1	*6420041*, *LitLink* ; Journal; Setola, Vincent; Hufeisen, Sandra J.; Grande-Allen, K. Jane; Vesely, Ivan; Glennon, Richard A.; Blough, Bruce; Rothman, Richard B.; Roth, Bryan L.; MOPMA3; Mol. Pharmacol.; EN; 63; 6; 2003; 1223 - 1229.

Pharmacological Data 2 of 57

Effect	receptor; binding activity
Species or Test-System	rat brain membranes
Concentration	0.01 - 10 µmol/l
Method	competition experiment: membranes incub. with title comp.+0.1 nmol/l <125I>DAIZAC at room temp.; bound and free radioligand separated by vacuum filtration, analysis by gamma spectrometry
Further Details	DAIZAC: (S)-5-chloro-3-iodo-2-methoxy-N-(1-azabicyclo<2.2.2>oct-3-yl)benzamide; title

Help, press F1 | idle | ALL | Substances

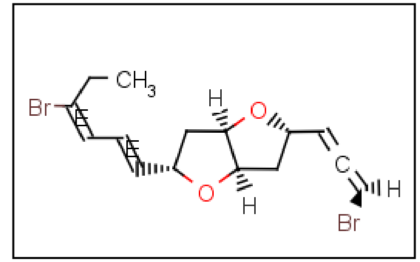
Trends in User Needs: Database Focus



- Beilstein originated as a dictionary of structures and their associated properties.
- One of the "properties" was the preparation (a reaction)
- In 2001, structures and reactions were equally popularly searched
- By 2007, structures searches tripled, but reaction searches increased x8

Aplysiallene Synthesis

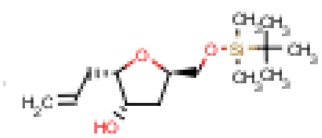
Wang, Jian; Pagenkopf, Brian L.
Organic Letters, **2007**, vol. 9, # 18 p. 3703 - 3706



First total synthesis and structural reassignment of (-)-aplysiallene

The first total synthesis of (-)-aplysiallene has been completed in 16 steps and features a key sequential Mukaiyama aerobic oxidative cyclization

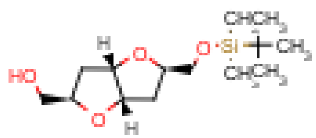
Related Structure	Referenced Compound
The configuration of the title compound previously suggested by: Suzuki, M.; Kurosawa, E. Phytochemistry 1985, 24, 1999-2002 is incorrect. The correct structure is given.	(-)-aplysiallene



7

82 %

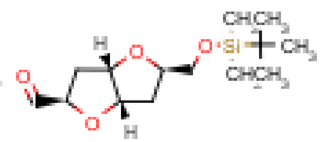
Add
Remove



6

86 %

Add
Remove

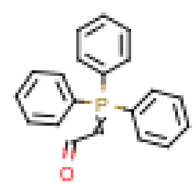


5

88 %

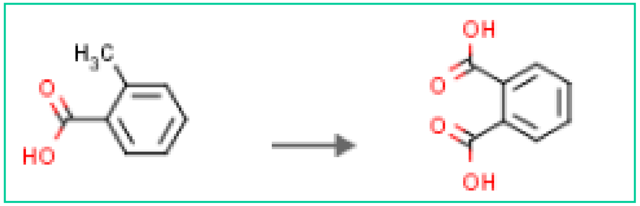
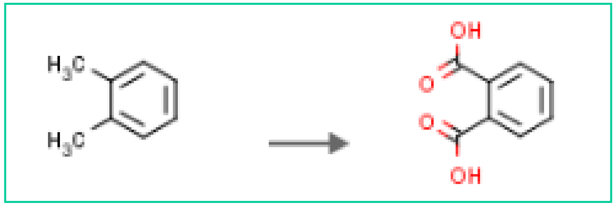
Add
Remove

7	82%	**With** tert.-butylhydroperoxide; oxygen; Co(II)(modp)2 **in** 2,2,4-trimethylpentane; isopropyl alcohol T=50°C; P=760.051 Torr; 16 h;



Synthesize

21st Century : Understanding Reaction Catalysis

Claus; Pieszcek | 1886 | Chemische Berichte, 1886 , vol. 19, p. 3084

21st Century:

Oxidation of substituted toluenes under phase-Transfer Catalysis conditions

Alkane Oxidation with Molecular Oxygen Using a New Efficient Catalytic System: N-Hydroxyphthalimide (NHPI) Combined with Co(acac)n (n = 2 or 3)

Oxidation of Alkylaromatic Hydrocarbons over V2O5-Sb2O3/TiO2 Catalyst

Solvent-Free Benzylic Oxidations Using Urea-Hydrogen Peroxide Complex (UHP) under Microwave Irradiation

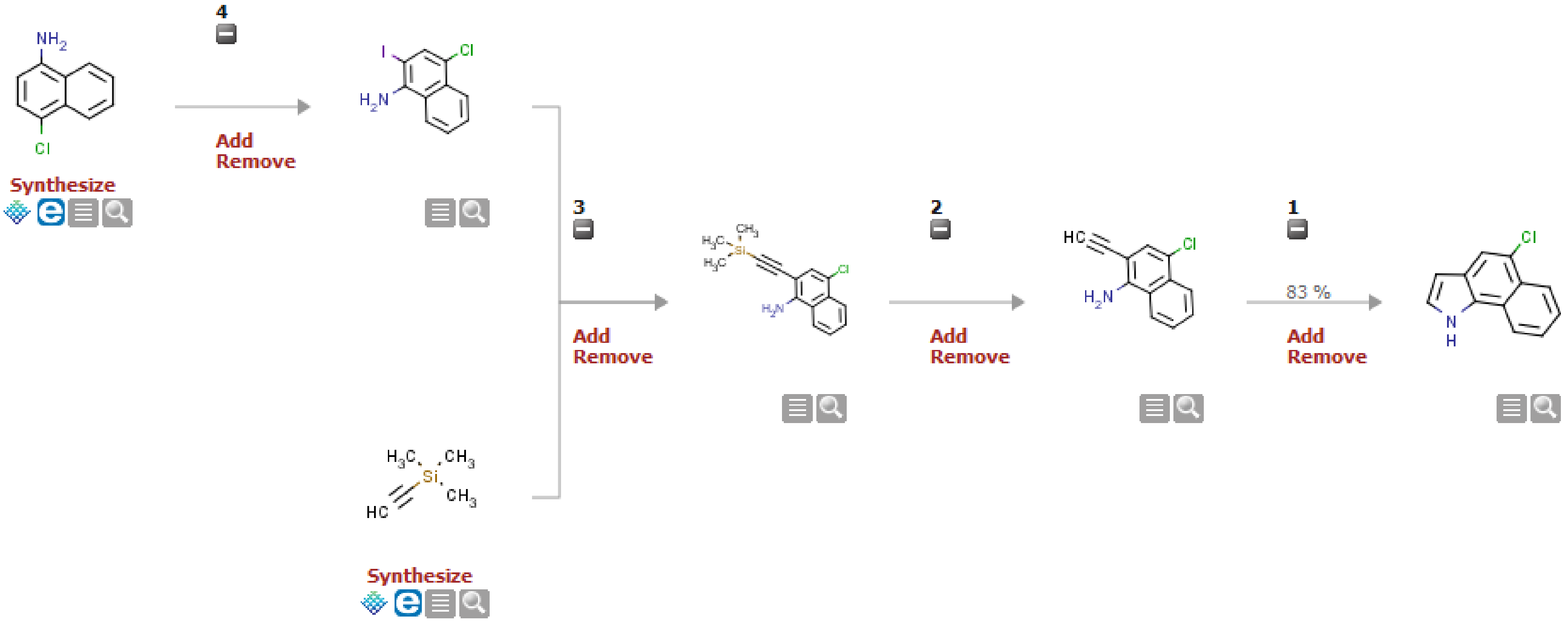
Is it Possible to Achieve Highly Selective Oxidations in Supercritical Water?
Aerobic Oxidation of Methylaromatic Compounds

ZnO: a versatile catalytic agent for benzylic oxidations

21st Century : Understanding Reaction Catalysis

Rhodium-Catalyzed Cycloisomerization: Formation of Indoles, Benzofurans, and Enol Lactones

Trost, Barry M.; McClory, Andrew
Angewandte Chemie, International Edition, **2007**, vol. 46, # 12 p. 2074 - 2077
Angewandte Chemie, **2007**, vol. 119, # 12 p. 2120 - 2123

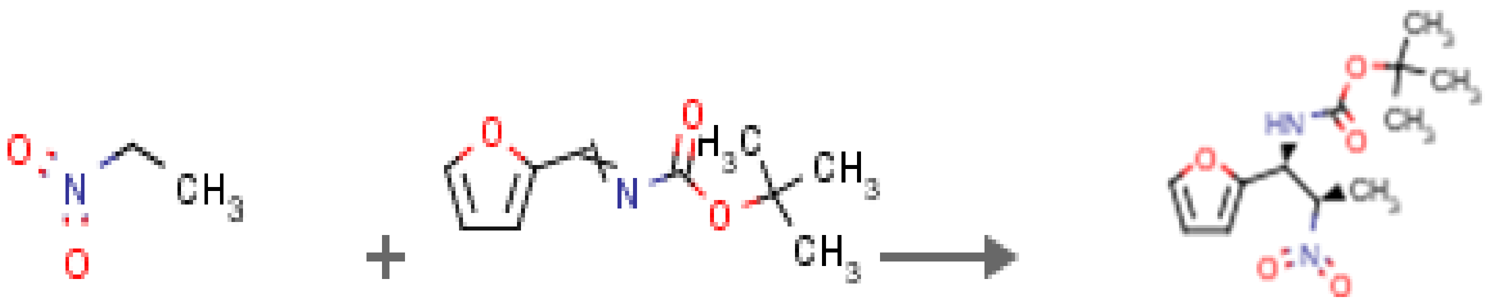


Multi-step reaction with 4 steps
1: benzyltrimethylammonium iodine dichloride; $CaCO_3$ / CH_2Cl_2; methanol / 25 °C
2: CuI; diethylamine / $PdCl_2(PPh_3)2$ / 25 °C
3: potassium carbonate / methanol / 25 °C
4: 83 percent / tris(4-fluorophenyl)phosphine / [Rh(COD)Cl]2 / dimethylformamide / 2 h / 85 °C

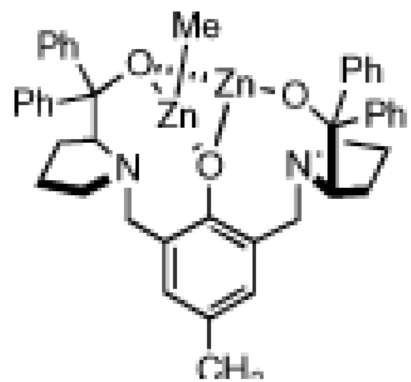
21st Century : Understanding Reaction Catalysis

Dinuclear Zinc-Catalyzed Enantioselective Aza-Henry Reaction

Trost, Barry M.; Lupton, David W.
Organic Letters, **2007**, vol. 9, # 10 p. 2023 - 2026



With prolinol-based zinc complex **in** tetrahydrofuran
T=-20°C; 14 h; aza-Henry reaction;



Conclusions and Open Questions

- Experimental data is the parent of workable models, and ultimately, knowledge.

- The most important data is probably that which does not fit the current model...
  - ...provided it is properly measured.
  - Do scientists need to publish data that **does** fit the current models?
  - And if so, are their peers prepared to pay for the considerable overhead costs of indexing and finding it?

- Some experimental data must always be able to be checked for validation, but has a shelf life in the shop windows of the secondary information services. We already see this trend in the provision of "Supplementary Data" in primary publication.
  - Who should store such supplementary data and make it available?

- Much of the historical background in this presentation was found on the internet, predominantly from Wikipedia.
  - I repeat: my profound thanks for that service.
  - But I would hesitate to say that I have acquired knowledge.
  - It is rather that I was capable of finding considered, evaluated responses to my queries.

  - Is that a model for consideration in the major databases?

Thank you for your attention...

- Any questions?