Some Common Misconceptions

1. The machines are here to replace us all.

2. All computational chemists are intimately familiar with informatics and machine learning

3. There are inherent limitations in machine learning methods that will prevent machines from ever designing syntheses of complex molecules (automated synthetic planning is “mission impossible”)

**What is Computer-Assisted Synthesis Planning?**

Computer assisted synthetic planning generally applies methods from informatics and machine learning to solve chemical problems.

<table>
<thead>
<tr>
<th>Starting Position</th>
<th>Relative Positions</th>
<th>End Position</th>
<th>Movements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predefined, with some moves not allowed</td>
<td>Discrete Configuration of Pieces</td>
<td>Checkmate or Draw</td>
<td>Standard Chess Moves (small, predefined number)</td>
</tr>
<tr>
<td>Random</td>
<td>Configuration of Cube</td>
<td>Successful Completion</td>
<td>Rotation of a layer</td>
</tr>
<tr>
<td>Synthetic Target</td>
<td>Synthetic Intermediates</td>
<td>Commercially available materials</td>
<td>Chemical Transformations</td>
</tr>
</tbody>
</table>

*If synthetic positions and synthetic moves can be defined, all of the above problems become very similar.*

G.E. Vléduts and V. K. Finn, 1957: an Information Machine for Chemistry will:

(i) search for individual chemical compounds ✓
(ii) search from chemical compounds possessing a certain given combination of characteristics ✓
(iii) search for classes of reactions into which a definite individual compound can enter ✓
(iv) search for the class of reaction producing a particular chemical compound ✓
(v) search for the class of reactions which are of the same type chemically and are characterized by a transfer of given structural elements... from the initial molecules into other definite structural elements of final molecules ✓
(vi) search for the reaction that will take place between given compounds under given conditions ✓
(vii) search for ways of synthesizing a given compound from a definite number of permissible initial compounds ?
Fifty Years of Development

DENDRAL Project, 1965 – Edward Feigenbaum, Bruce Buchanan, Joshua Lederberg, and Carl Djerassi

Organic Chemical Simulation of Synthesis (OCSS), 1969 – Corey and Wipke

Logic and Heuristics Applied to Synthetic Analysis (LHASA) – Corey and Wipke

Simulation and Evaluation of Chemical Synthesis (SECS) – Wipke

SYNCHM 1977-1998 – Stanford/Stony Brook

SYNLMA, 1989 – P. Y. Johnson

SYNGEN, 1977-1990 – J. B. Hendrickson

IGOR/IGOR2, 1974-1993 – Ugi

CHIRON, 1990-2005 – Stephen Hanessian

WODCA - Johann Gasteiger

ARChem Route Designer – SymbioSys

ICSYNTH – ChemInfo

Chematica – Gryzbowski

Bartosz Gryzbowski:
PhD. Harvard 2000 (Whitesides)
Post-doc Harvard 2000-2003
Northwestern (2003-2014)
UNIST (Distinguished Professor, Chemistry, 2014~Present)
ProChimia Surfaces (Chief Scientific Officer, 2002-Present)
GSI L.L.C. (President, 2009-Present)
Position and Moves

Logical Connections

Nonsensical Connections

Position and Moves


The Network of Organic Chemistry

NOC acquired from Beilstein Database
Over ten million unique structures as SMILES/SMARTS notation with ten million connections

Scale-Free Architecture

Is it possible to navigate the NOC rapidly generate synthetic pathways of known targets?

Transversing the NOC

Breadth First vs. Depth First Searches

“The combinatorial explosion”
-E.J. Corey

Movie Break
Search Criteria and Constraints

Cost:

\[ C_{tot} = C_{rxn}N_{rxn} + \sum_i C_{sub}(i) \]

Popularity:

Function of \( k_{in} / k_{out} \)
Example: Synthesis of Gabapentin

Red Nodes Denote Commercially Available Materials

Blue Nodes Denote Synthetic Intermediates

Yellow Halos Denote Controlled Substances

Golden Node Denotes Target Compound
Example: Synthesis of Zolpidem

\[ C_{rxn}^o = 7.5 \]

\[ C_{rxn}^o = 0.0075 \]

$2.76 / g$

$196 / mol$

$1.61 / g$

$175 / mol$

\[ \text{Peptide Coupling} \]

\[ \text{CuCl (5 mol %)} \]

\[ \text{CuOTf}_2 (5 \text{ mol %)} \]

toluene, \(120\) °C

\(72\%\)

Angew. Chem. Int. Ed. 2010, 49, 2743 – 2746

Example: Synthesis of Vardenafil
Development of One-Pot Reactions

Assign Molecules by Functional Groups

Compare Classification Against a 322 x 322 master grid of functional group compatibility

If incompatible groups exist, either: 1) suggest compatible order of addition or 2) omit from candidate combinations

86,000 Chemical Criteria

Over 1 million 2-step sequences

14 two-step, 12 three-step, one 4-step sequences experimentally evaluated

1. Assign Molecules by Functional Groups

2. Compare Classification Against a 322 x 322 master grid of functional group compatibility

3. Acid / Base Compatibility

4. Solvent Miscibility

5. Hydride / Proton Incompatibility

6. Aqueous vs Nonaqueous

7. Compatibility between Reagents and Functional Groups

8. Oxidizing vs Reducing

Assess if all functional groups are compatible with all reaction conditions

Selected Examples of One-Pot Reactions

Nothing's Perfect...

“Intelligent” Retrosynthetic Analysis

Navigating the NOC might make the computer seem smart... but is it really?

1) In NOC, the only positions and moves available are taken directly from the literature – novel transformations or novel compounds are unattainable

2) In NOC, all synthetic positions are static – Expert organic chemists use a dynamic network

How do we make a computer “smart enough” to solve synthetic problems?

Fifty Years of Development

- DENDRAL Project, 1965 – Edward Feigenbaum, Bruce Buchanan, Joshua Lederman, and Carl Djerassi
- WODCA - Johann Gasteiger
- Organic Chemical Simulation of Synthesis (OCSS), 1969 – Corey and Wipke
- Logic and Heuristics Applied to Synthetic Analysis (LHASA) – Corey and Wipke
- Simulation and Evaluation of Chemical Synthesis (SECS) – Wipke
- SYNLMA, 1989 – P. Y. Johnson
- SYNGEN, 1977-1990 – J. B. Hendrickson
- IGOR/IGOR2, 1974-1993 – Ugi
- CHIRON, 1990-2005 – Stephen Hanessian
- ARChem Route Designer – SymBioSys
- IC~3~SYNTH – Cheminfo
- Chematica – Gryzbowski

Possible Reasons for failure:

- Needed Better Computers / Algorithms
- Oversimplification of the Problem

Exhaustive Searches vs. Heuristics

How do we do exhaustive searches on dynamic networks? How does this influence synthetic position? How do we account for sparse events / stereochemistry / molecular context?
Syntaurus

Molecular Representations (SMARTS/SMILES) combined with a set of expert rules

Over 20,000 rules (in 2016, Aldrich website says 50,000), 200,000 specialized reactions in addition to “conditional rules of chemistry”
Early Failures

Attempt 1: Use machine-extracted transforms as rules
– ca. 115,000 unique reaction classes

Errors in Databases

“Context dependent” Cases

Non-synthetically Useful Conditions

Relative Abundance of Products

Group Problem 1

Aripiprazole
Built-in script language for custom scoring:

- **MREL**: favorably scores substrates of comparable molecular weights
- **STEREO**: favorably scores creating enriched stereocenters
- **RINGS**: favorably scores creating rings
- **BUY**: promotes substrates that are commercially available
- **CONFLICT**: Penalizes reactions with functional group incompatibility
- **PROTECT**: Penalizes the necessity for protecting groups
Automation

Individual searches are inefficient – manual selection of each step makes finding the optimal route highly improbable and very time consuming.

Searching on a Dynamic Network is more difficult – no information is available about subsequent layers and evaluation of step and position is necessary.

Chemical Scoring Function
RINGS, STEREO, MASS, SMILES_LEN, KNOWN, WEIRD, BUY, KNOWN

Reaction Scoring Function
PROTECT, CONFLICT, YIELD

If CSF = 0 and RSF = 1:
Each reaction step costs +1 – minimizes number of steps.

Estimation of Reaction Yields

Concept: Calculate accurate $G^\text{form}$ and use them to calculate $\Delta G^\text{rxn}$, which should then correlate with yield.

Assumptions: Most reactions are under thermodynamic control, the training set is representative of most chemical reactions.

1) Obtain Training Set of 23,000 reported reactions (MW of reactant 100-1000 g / mol)

2) Decompose training molecules into 296 distinct functional groups and assign guess $\Delta G^\text{form}$ values

3) $\Sigma \Delta G^\text{form} = \Delta G^\text{calc}$

$\Delta G^\text{calc}$ does not account for non-ideality.

Must use experimental yields with perturbed-chain statistical associating fluid theory to attain more accurate values.
Estimation of Reaction Yields

4a) Solve for mole fraction

\[ \xi = \frac{n_i^o - x_i n^o}{x_i \nu - \nu_i} \]

\( \xi \) = Experimental Yield

\( n_i^o \) = Initial mols of \( i \)

\( x_i \) = mol fraction \( i \)

\( n^o \) = total number of initial mols

\( \nu \) = total stoichiometry coefficient

\( \nu_i \) = stoichiometry coefficient for \( i \)

4b) Use PC-SAFT to calculate \( \gamma \)

\[ \Delta G^{exp} = -RT \ln \prod (x_i \gamma_i)^{\nu_i} \]

5) Optimize \( \Delta G^{form} \) to fit experimental data

\[ OBJ = \sqrt{(\Delta G^{exp} - \Delta G^{calc})^2/2} \]

Approx. 15% error – still good enough to provide qualitative assessment.

Seminal publication of a this approach to predict reaction yield.

Angew. Chem. Int. Ed. 2015, 54, 10797 – 10801
Selected Examples of Yield Prediction

(-)-7-Methylmuralide (Corey and Shenvi)

\[
\text{Ketone} + \text{Phosphoramidite} \xrightarrow{\text{CH}_2\text{Cl}_2, \text{NEt}_3} \text{Product}
\]

Predicted: 90%  
Observed: 75%

Marinopyrrole A (Nicolaou)

\[
\text{Imide} + \text{Phosphoric Acid} \xrightarrow{\text{CH}_2\text{Cl}_2} \text{Product}
\]

Predicted: 59%  
Observed: 64%

<table>
<thead>
<tr>
<th>Solvent</th>
<th>Yield (P/O)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMF</td>
<td>91% / 82%</td>
</tr>
<tr>
<td>DMA</td>
<td>43% / 59%</td>
</tr>
<tr>
<td>DMPU</td>
<td>0% / 12%</td>
</tr>
</tbody>
</table>
Searching in Syntaurus

Search algorithm should be 1) non-local 2) strategizing and 3) self-correcting
Movie Break
“Rediscovery” of Published Synthesis

Tantillo, Aubé, et al. J. Org. Chem. 2015, 80, 5260−5271
Rediscovery of Published Synthesis
Rediscovery of Published Synthesis


E: Exact Procedure

Taylor et al, Org. Lett. 2013, 15, 258-261
Group Problem 2

tacamonidine
goniothalesdiol A
juvabione
(R)-6-((2S,4S)-2,4-dihydroxypentyl)-5,6-dihydro-2H-pyran-2-one
Group Problem 2

A

B

C

D

E

Group Problem 2

A) **Straight from Literature**

B) 

C) 

Group Problem 2

**Group Problem 2**

![Reaction Scheme]

**b)**

1. **Step a**
   - Reagents: CH(SEt)_3, n-BuLi, THF-HMPA (5:1)
   - Temperature: 96 %
   - Reaction: CH(SEt)_3, n-BuLi in THF-HMPA (5:1)
   - Product: \( \text{C(SEt)}_3 \)

2. **Step b**
   - Reagents: n-BuLi, BF_3 OEt_2, DCM
   - Temperature: -78 °C
   - Product: \( \text{TMS} \)

3. **Step c**
   - Reagents: DIBAL-H, CuCl_2, CuO
   - Temperature: MeOH-H_2O-CH_2Cl_2 (10:1:1)
   - Reaction: CuCl_2, CuO in MeOH-H_2O-CH_2Cl_2 (10:1:1)
   - Product: \( \text{CO}_2\text{Me} \)

4. **Step d**
   - Reagents: PhCHO, InBr_3, CH_2Cl_2, r.t., hr
   - Temperature: 72 %
   - Product: \( \text{I} \)

5. **Step e**
   - Reagents: NMO, OsO_4 (cat), THF-H_2O (3:1)
   - Temperature: 0 °C to r.t.
   - Product: \( \text{OH} \)

References:
She, Synlett 2010, 15, 2283 – 2284
Group Problem 2

Yamamoto, JACS. 2007, 129, 2762-2763

\[
\begin{align*}
\text{(Me}_3\text{Si)}_3\text{Si} & \quad + \quad \text{O} \quad \text{OTBS} \\
\text{HNTf}_2 \quad \text{(0.05 mol %)} & \quad \longrightarrow \quad \text{MgBr} \quad \longrightarrow \quad \text{Cl} \quad \text{O} \quad \text{OTBS} \\
\text{-78 °C, 30 min} & \quad \text{-90 °C, 30 min} \quad \text{-90 °C, 30 min} \\
\text{63 %} & \\
\text{Grubbs' G2} & \quad \longrightarrow \quad \text{O} \quad \text{OTBS} \\
\text{88 %} \quad \text{then acetic anhydride} & \quad \longrightarrow \quad \text{OAc} \quad \text{OAc} \\
\text{75 % (32 % overall yield)} & 
\end{align*}
\]
Points of Improvement

More efficient searches: intrinsic (molecular topology) vs extrinsic (number of stereocenters created) metrics, synthetic accessibility, outcomes dictated by non-local contributions, the combinatorial explosion

Reaction rules: Back to machine-learned rules?

Cao and Liu’s Topological Steric Effect Index
Conformer Distributions?
Quantum Chemical Calculations?
A Machine that thinks like a Chemist!

Chematica is an unprecedented decision-making and synthetic planning software product. This expert system combines an incredible amount of chemical knowledge and processes it in intelligent ways within seconds. In addition to its unrivaled speed, its ability to design viable and optimized synthetic pathways towards both known and previously-unexplored targets remains unmatched. Today, Chematica is the indispensable companion of the 21st century chemist and the new catalyst for the everyday practice of organic synthesis and chemical discovery.

The world-wide press has hailed Chematica as paradigm shifting and dubbed it:

“Automatic Chemist”… by Philip Ball in Chemistry World

“Chemical Internet”… by Ian Tucker in the Guardian

“Robo-Chemist”… by Mark Peplow in Nature

“Immortal Chemist”… by Daily Mail Reporter