

Computer-Aided Synthesis Planning: a brief history



Shutian Jiang
Literature presentation
Jul. 28, 2022

Outline

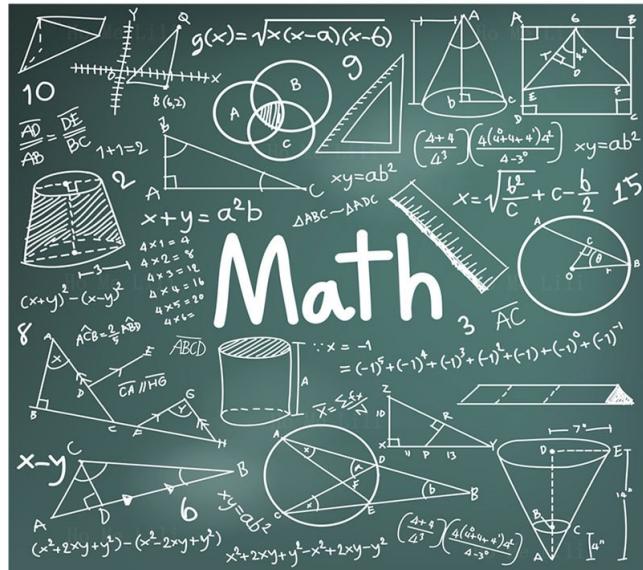
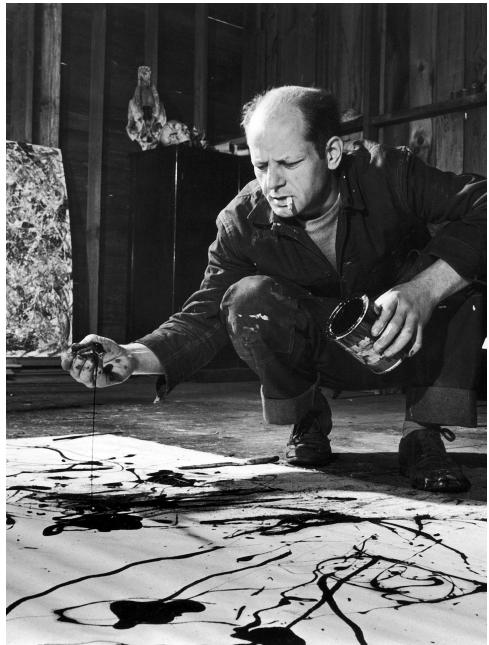
- Introduction: what is total synthesis
- History of E. J. Corey and LHASA
- Breaking down the different aspects:
 - Retrosynthetic analysis
 - Reactivity prediction
- Case Studies (Cernak & Newhouse)
- Outlook and Conclusion

What is total synthesis?

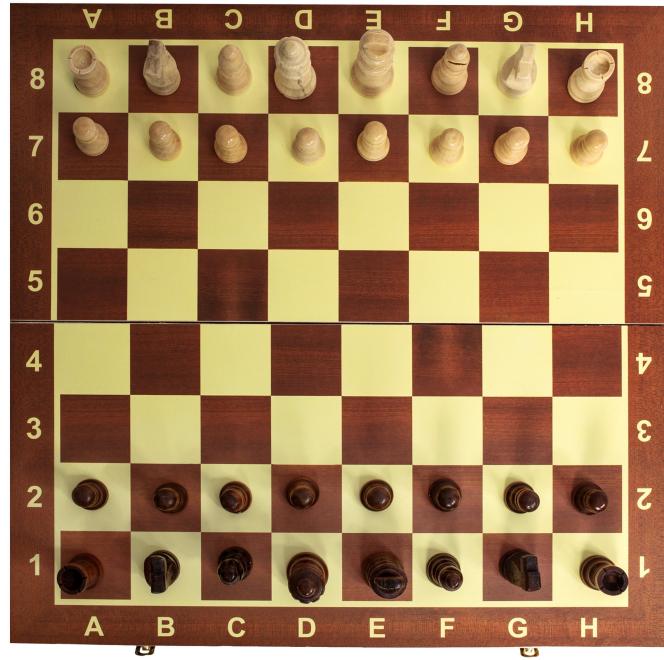
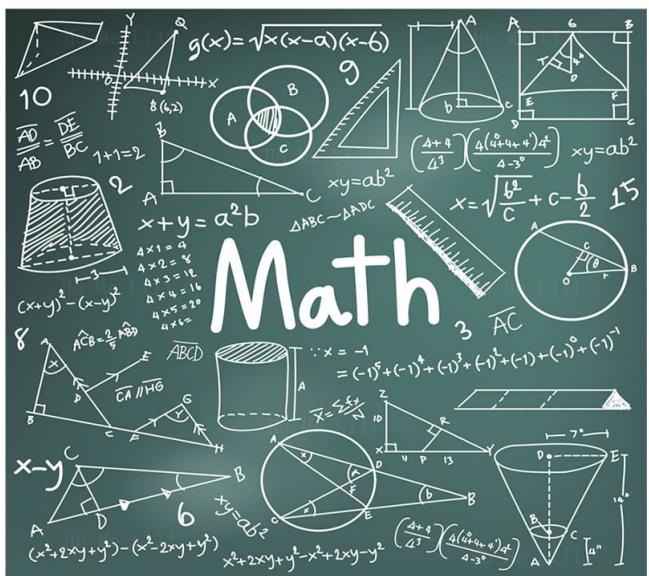
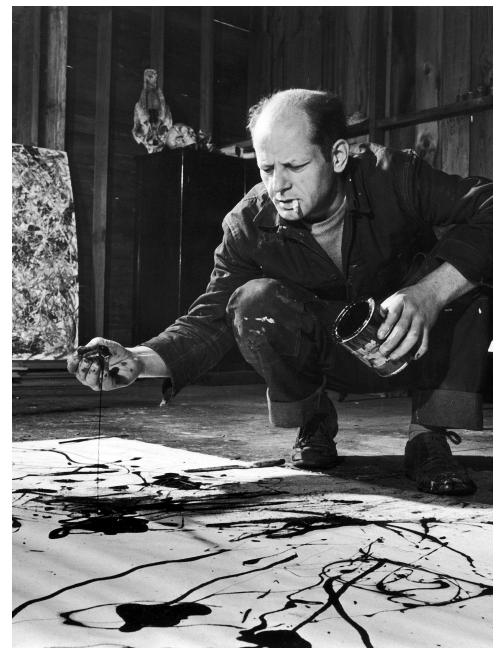
What is total synthesis?



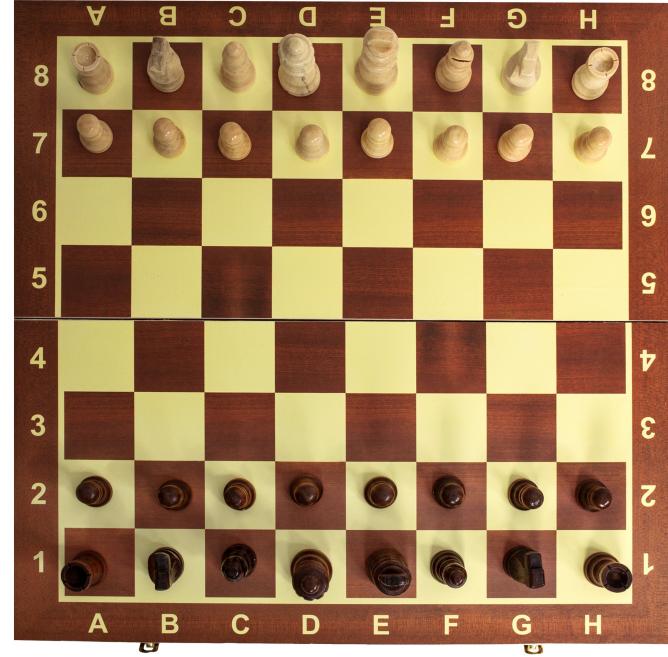
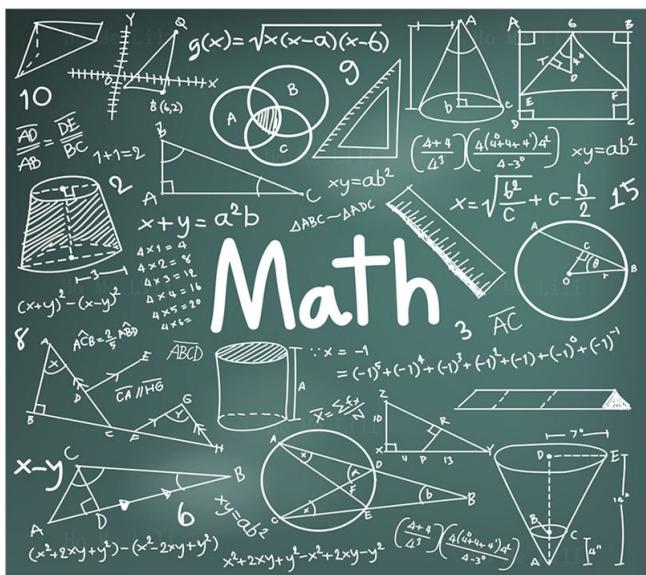
What is total synthesis?



What is total synthesis?

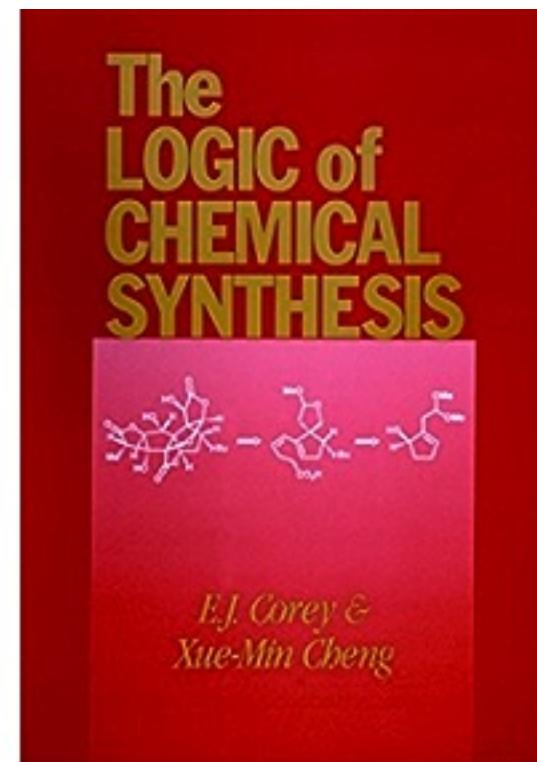


What is total synthesis?



- How do we teach a computer to do total synthesis?
- How do synthetic chemists benefit from the use of computers?

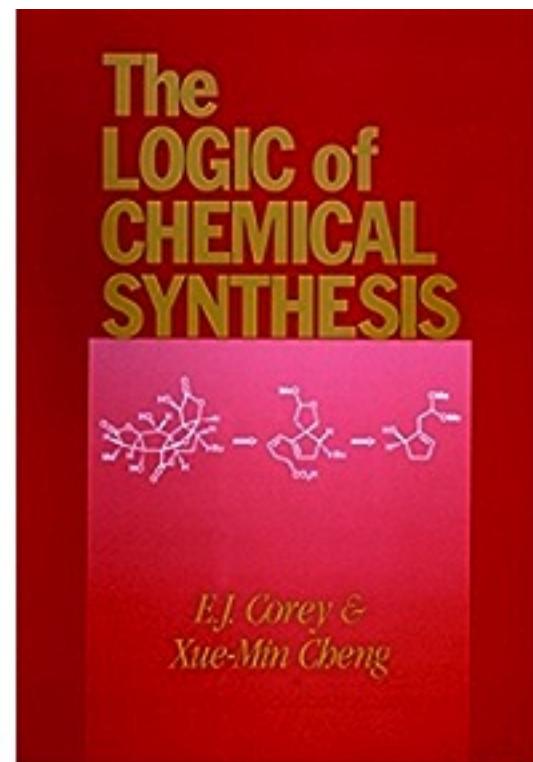
Where it all begins: E. J. Corey and LHASA



Pensak, D. A.; Corey, E. J. *Computer Assisted Organic Synthesis* 1977, Ch. 1, pp 1-32.

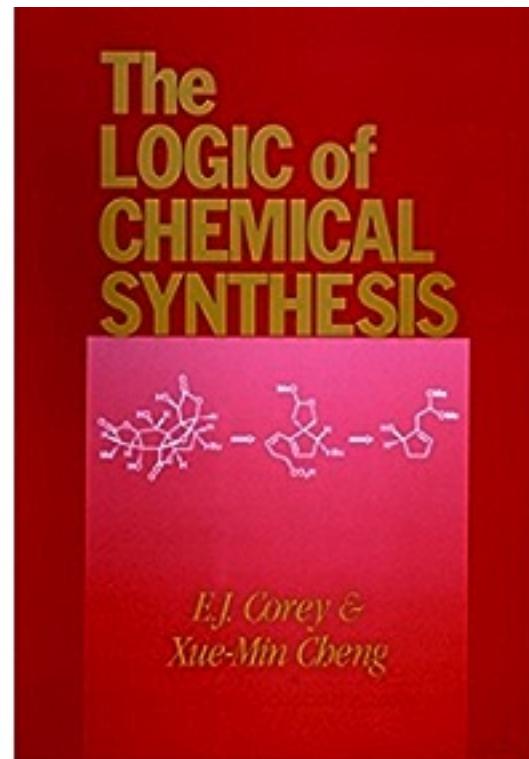
Where it all begins: E. J. Corey and LHASA

- What is LHASA?
- Logic and Heuristics Applied to Synthetic Analysis
- The goal of LHASA: design a “general purpose computer program” to aid chemists with synthetic routes employing “both the basic and more complex techniques”

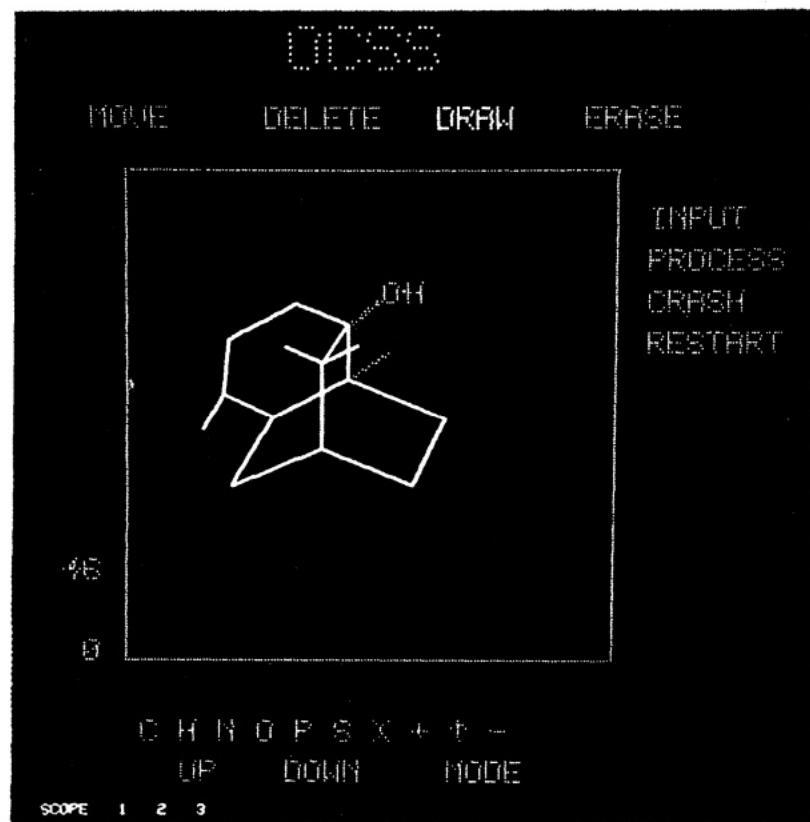
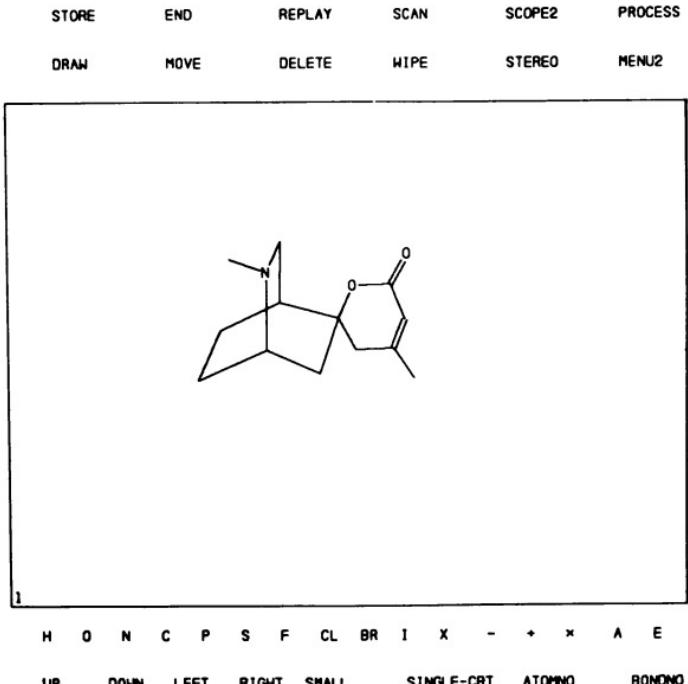


Breaking down LHASA:

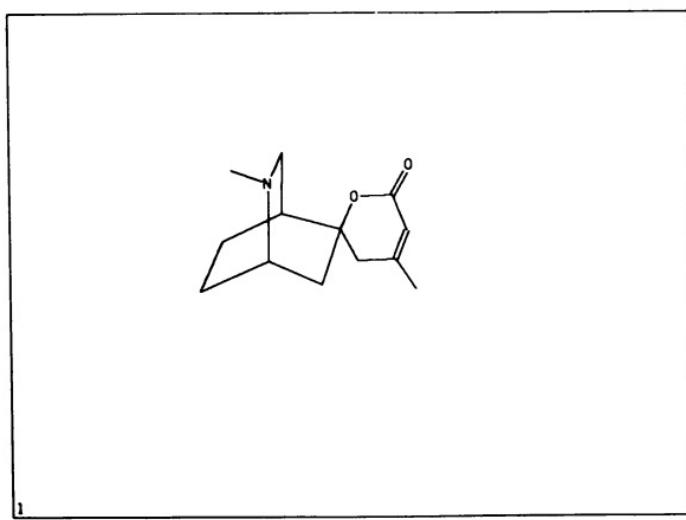
- Input/reading of molecular structures
- Strategy and control:
 - Fundamental strategies
 - High-level strategies
- Structure of each module: an example of hand-encoded rules of reaction



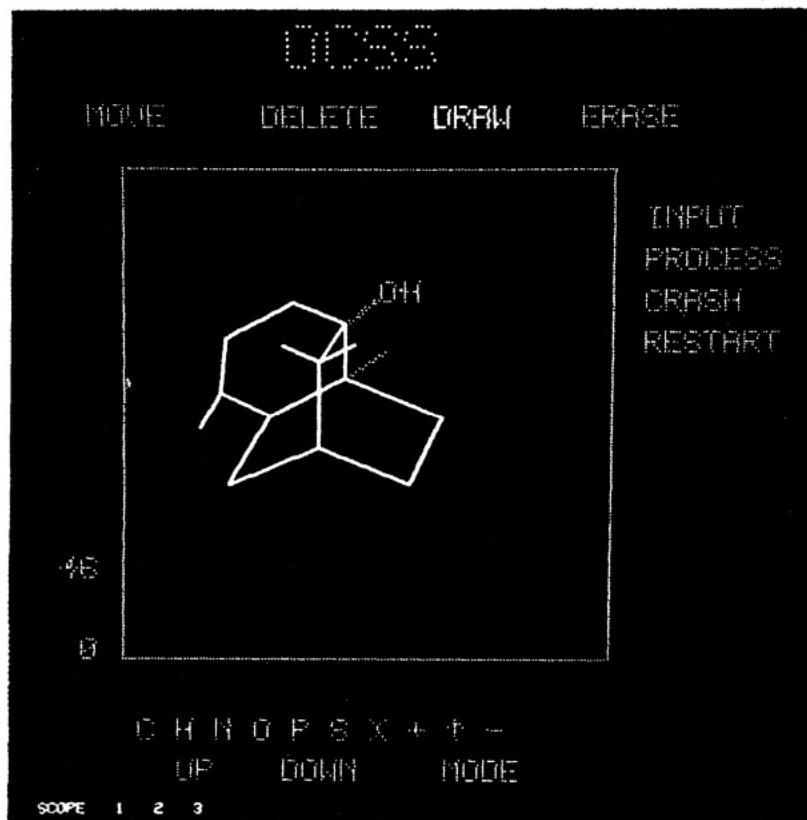
Input/reading of molecular structures



Input/reading of molecular structures

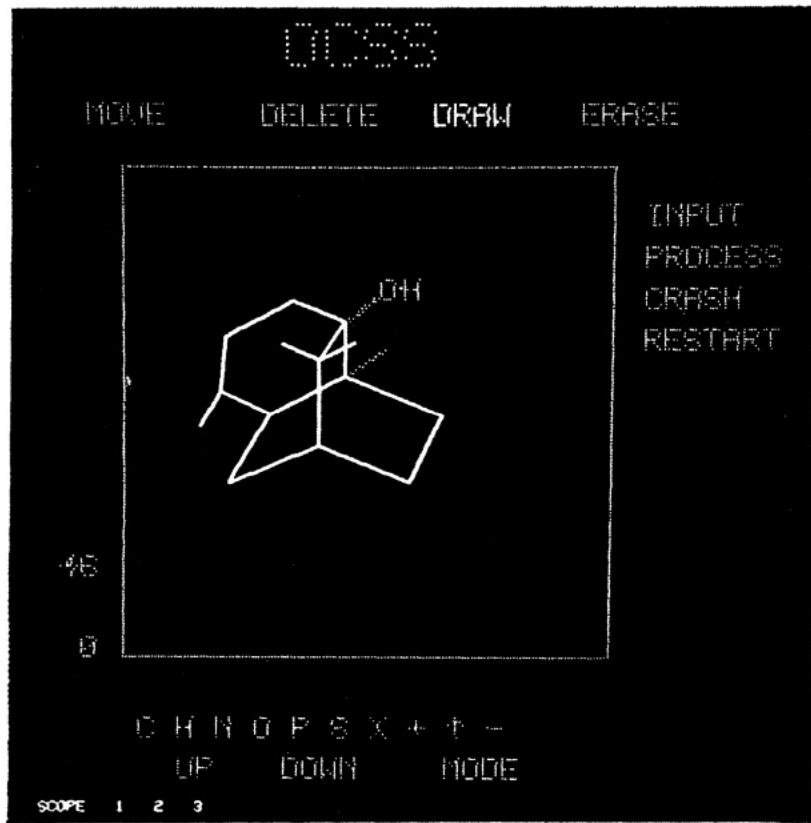
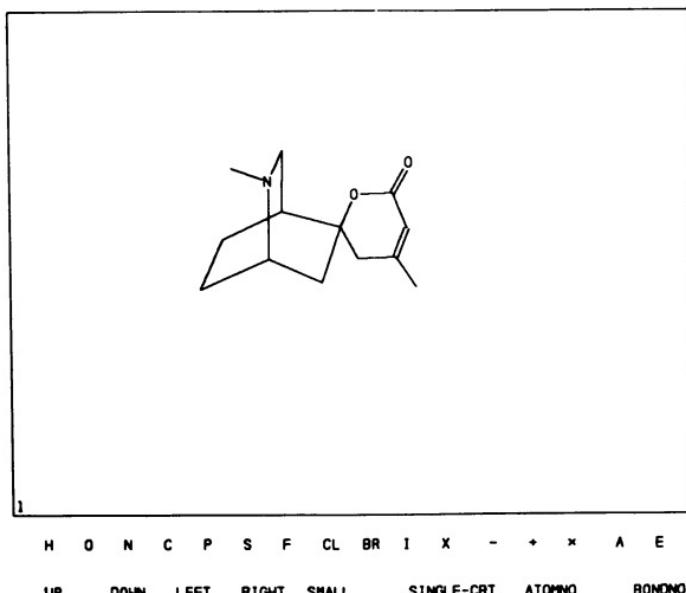


H O N C P S F CL BR I X - + X A E
UP DOWN LEFT RIGHT SMALL SINGLE-CRT ATOMNO BONDNO



- Drawn with stylus on a tablet; displayed on a CRT screen

Input/reading of molecular structures



- Position, type, connectivity of the atom/bond
- Chemical knowledge: Functional groups and significant rings:
 - Identify unstable arrangements/tautomers
 - Classify stereocenters and double bonds etc.

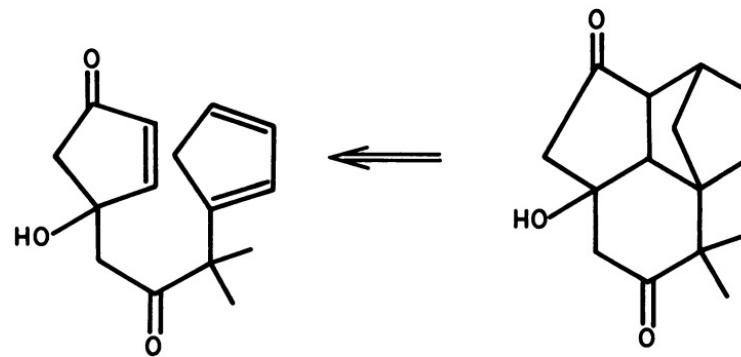
Perception of rings and functional groups

- Perceptions of rings:

- Identification based solely on connectivity (not orientation)
- Real and Pseudo rings

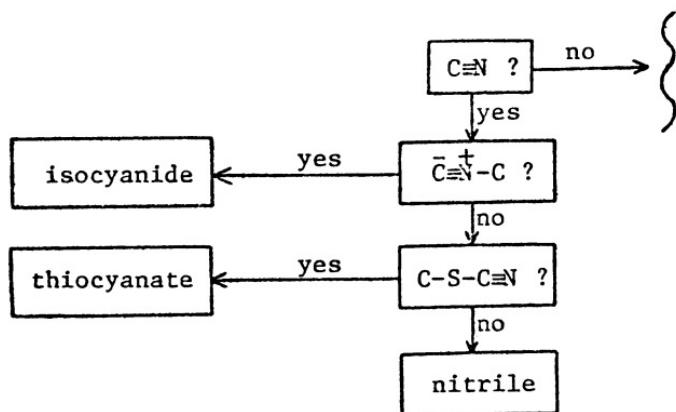


- Strategic bond disconnections in cyclic structures



Perception of rings and functional groups

- Perceptions of functional groups:
 - Identification of “juxtapositions” of groups
 - Recognition of “origin” atoms and reactive sites
 - Manually-encoded rules for identification



A22	LOC A23	LOC A24	SHIFT + IF CARBON*COUNT IS TWO
A23	NULL	NULL	IDENTIFIED AS KETONE
A24	LOC A25	LOC A26	IF HYDROGEN*COUNT IS TWO
A25	NULL	NULL	IDENTIFIED AS ALDEHYDE
A26	LOC A27	LOC A28	IF HYDROGEN*COUNT IS ONE
A27	LOC A25	LOC A28	IF CARBON*COUNT IS ONE
A28	LOC A29	LOC A32	SEARCH FOR C**N
A29	LOC A30	NULL	SHIFT + SEARCH FOR C*N
A30	LOC A31	LOC A31	NONORIGIN ENTRY
A31	NULL	NULL	IDENTIFIED AS ISOCYANATE
A32	LOC A33	LOC A33	ENTRY BOND*SHARED
A33	LOC A34	LOC A43	SEARCH FOR C*O
A34	LOC A35	LOC A35	BOND*SHARED
A35	LOC A36	LOC A37	SHIFT + IF HYDROGEN*COUNT IS ONE
A36	NULL	NULL	IDENTIFIED AS ACID
A37	LOC A38	NULL	SEARCH FOR C*O
A38	LOC A39	LOC A39	SHIFT + NONORIGIN
A39	LOC A40	LOC A40	BOND*SHARED
A40	LOC A41	LOC A42	SHIFT + IF IN RING OF ANY SIZE
A41	NULL	NULL	IDENTIFIED AS LACTONE
A42	NULL	NULL	IDENTIFIED AS ESTER
A43	LOC A44	LOC A45	SEARCH FOR C*X
A44	NULL	NULL	IDENTIFIED AS ACID*HALIDE
A45	LOC A46	LOC A66	SEARCH FOR C*N
A46	LOC A47	LOC A47	BOND*SHARED
A47	LOC A48	LOC A49	SHIFT + IF HYDROGEN*COUNT IS TWO
A48	NULL	NULL	IDENTIFIED AS AMIDE*1

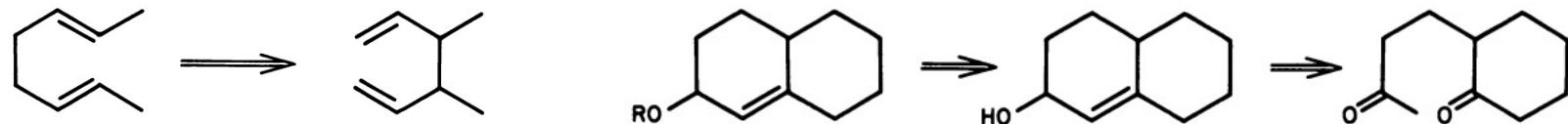
Selection of the strategies

- Broad categories of synthetic strategies:
 - Functional group-based transforms
 - Strategic bond disconnections for polycyclic targets
- Structural features based:
 - Appendages
 - Ring
 - Masked functionality

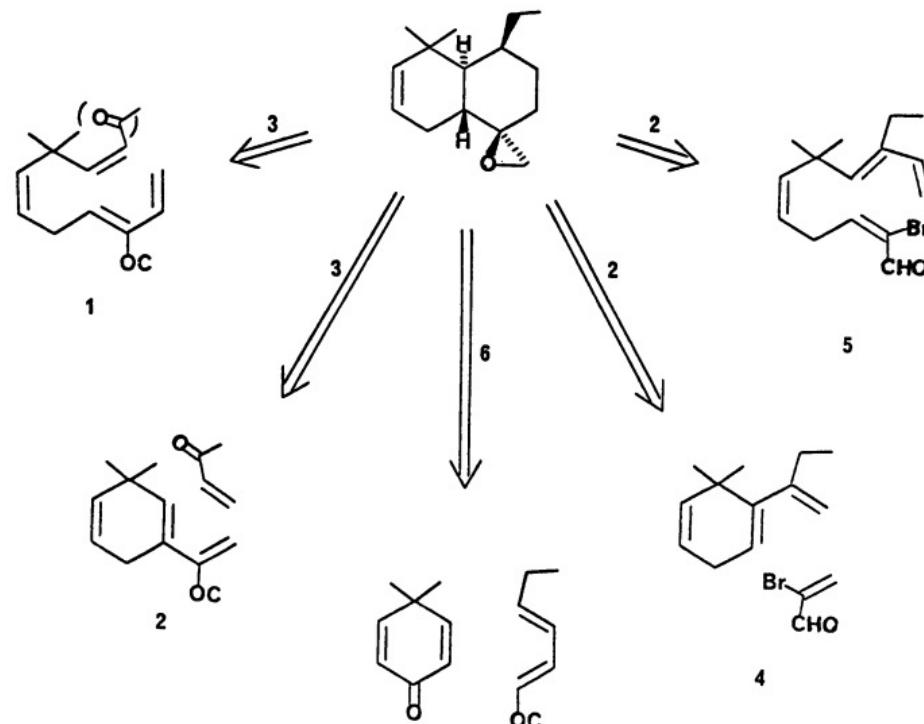
Selection of the strategies

- Broad categories of synthetic strategies:

- Functional group-based transforms



- Strategic bond disconnections for polycyclic targets



- Structural features based:
 - Appendages
 - Ring
 - Masked functionality

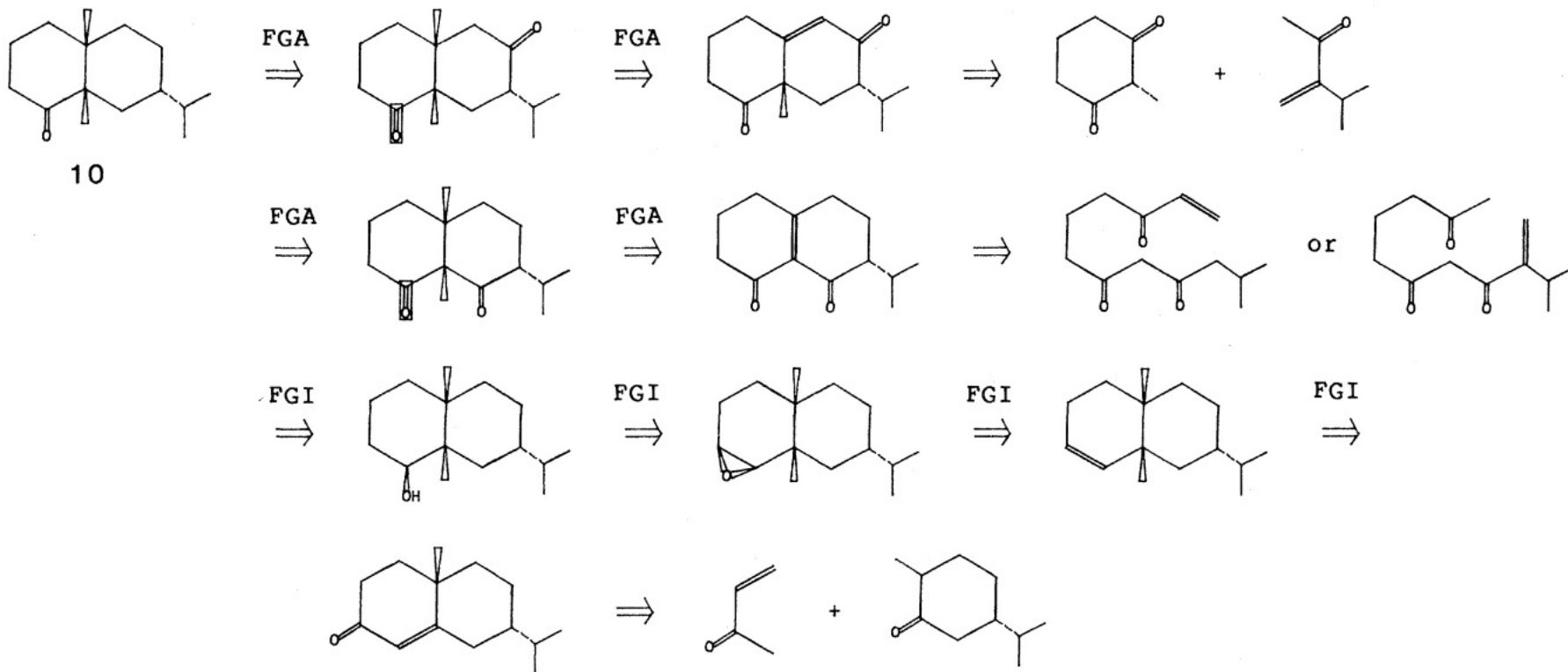
Example for encoded transformations

```
TRANSFORM 117
NAME MICHAEL ADDITION OF HETERO NUCLEOPHILE
...HET-C-C-W => HET-H + C=C-W
    ...MARCH 585; HOUSE 596; B+P 468
    ...ORG. RXNS. VOL.5, 79-135 (1949)
    ...BULL. SOC. CHEM. FR. 254,325 (1962)
    ...PATH 2 BONDS
RATING 50      ...Old rating 40
GROUP*1 MUST BE KETONE OR CYANO OR ESTER OR ACID
    OR LACTONE OR AMIDE*3 OR AMIDE*2 OR AMIDE*1
    OR LACTAM OR VINYLW OR ALDEHYDE
GROUP*2 MUST BE ETHER OR AMINE*1 OR AMINE*2 OR AMINE*3
    OR SULFIDE OR THIOL
STUDENT
REMOVES*STEREO CARBON2*1 ATOM*2
BROKEN*BONDS BOND2*1

...
KILL IF NO HYDROGEN ON ATOM*2
    ...REQUIRED FOR REACTION
KILL IF MULTIPLE BOND ON ATOM*2 OFFPATH OR: ON ATOM*3 OFFPATH
    ...WOULD PRODUCE ALLENIC PRECURSOR
IF BOND2*1 IS NOT IN A RING OF SIZE 5 THROUGH 7  &
    THEN KILL IF BOND2*1 IS IN A RING
SUBTRACT 15 IF LEAVING GROUP ON ATOM*3
    ...POSSIBLE ELIMINATION
ADD 15 IF ANOTHER WITHDRAWING BOND ON ATOM*2
    ...EASIER ADDITION
SUBTRACT 15 FOR EACH WITHDRAWING BOND ON ATOM*3
    ...UNDESIRED MICHAEL POSSIBLE
SUBTRACT 10 IF ATOM*3 IS A TERTIARY*CENTER
IF NOT OLEFIN ON BOND*2 THEN KILL IF ATOM*2 IS NOT ENOLIZABLE
    ...STABLE ENOL PROVIDES DRIVING FORCE
IF SECOND GROUP IS ETHER THEN CONDITIONS NaOR
IF SECOND GROUP IS AMINE THEN CONDITIONS RNH2
IF SECOND GROUP IS SULFIDE OR: THIOL THEN  &
    CONDITIONS NaSR
...
    BREAK BOND2*1
    JOIN ATOM*2 AND ATOM*3
...
```

- Each reaction (with compatible FG's etc.) hand-coded as above

Sample output sequence



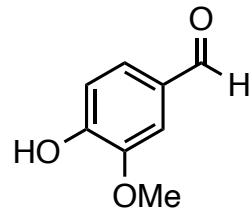
- Robinson annulation module; retrosynthesis of valerenone

Room for improvements

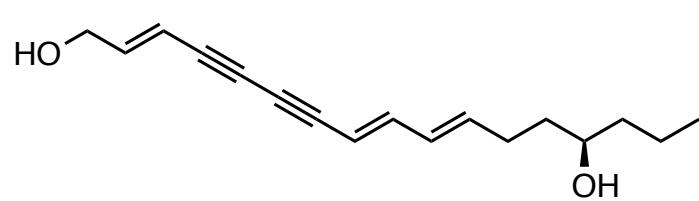
- Structural input and processing
- Chemist input: selection of module; choice of endpoint
- Evaluation/scoring of routes
- Database of reactions/templates: need to be manually added

New forms of machine-readable structure information

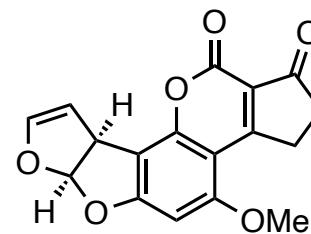
- Smiles strings: simplified molecular-input line-entry system
 - Also human readable (sort-of)
 - Rapid extraction & generation of molecular features
 - Substructure search; similarity parameters; molecular fingerprints



vanillin
COc1cc(C=O)ccc1O



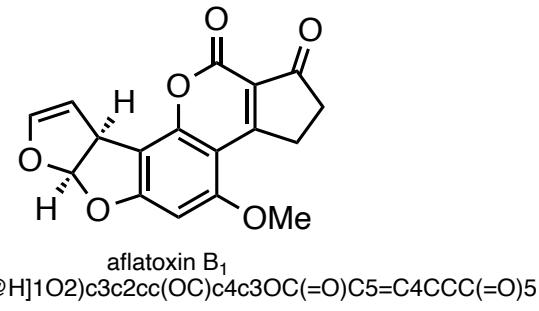
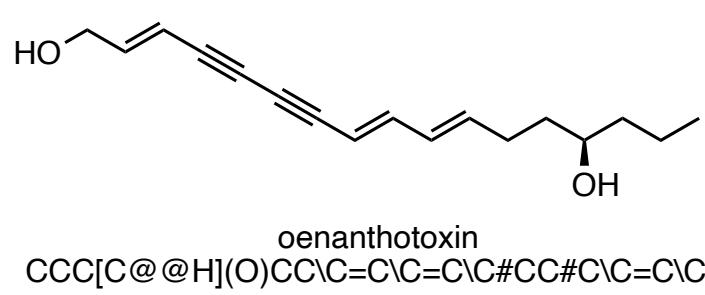
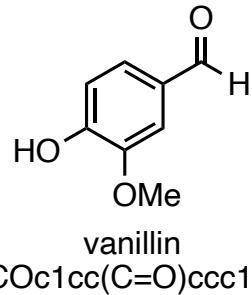
oenanthotoxin
CCC[C@H](O)CC\C=C\C=C\C#CC#C\C=C\CO



aflatoxin B₁
O1C=C[C@H]([C@H]1O2)c3c2cc(OC)c4c3OC(=O)C5=C4CCC(=O)5

New forms of machine-readable structure information

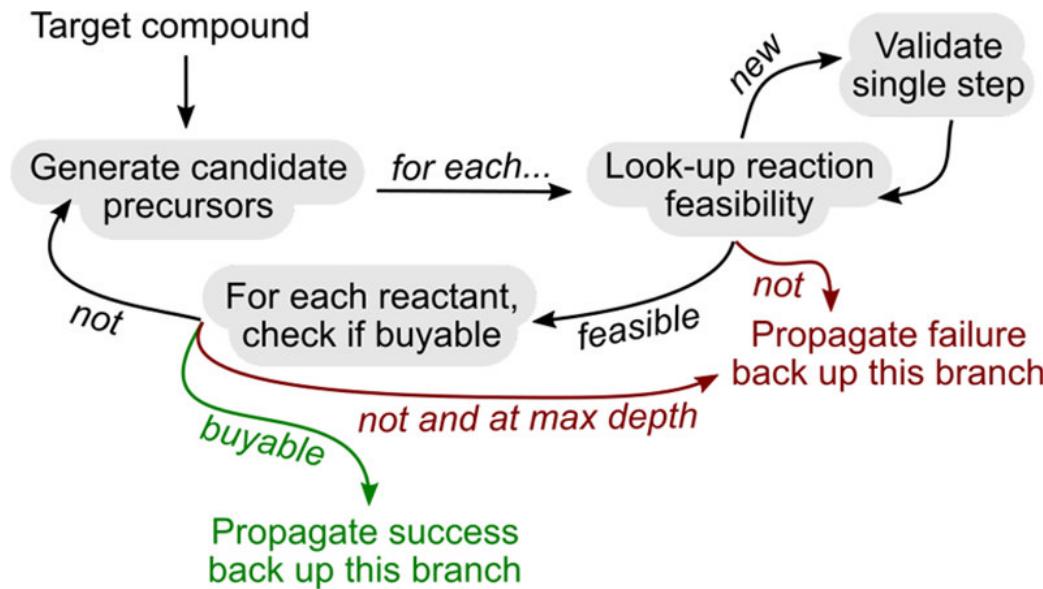
- Smiles strings: simplified molecular-input line-entry system
 - Also human readable (sort-of)
 - Rapid extraction & generation of molecular features
 - Substructure search; similarity parameters; molecular fingerprints



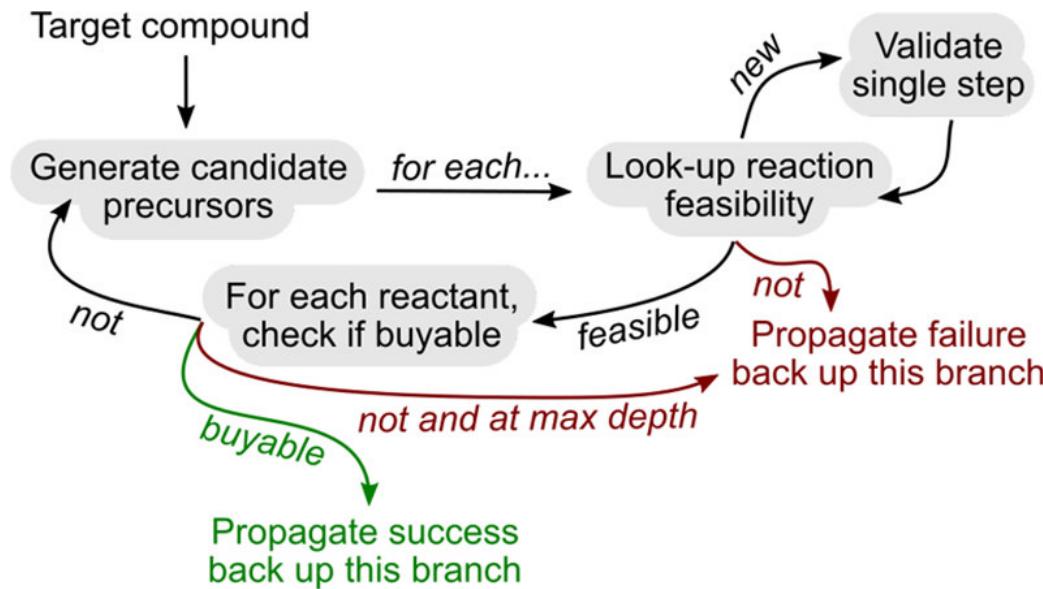
- InChI: international chemical identifier (IUPAC)
 - Less human readable
 - Unique to each molecule
 - InChIKey: fixed length (to truncate long strings from large structures)

EtOH: InChI=1S/C2H6O/c1-2-3/h3H,2H2,1H3

Similarities and differences of present algorithms



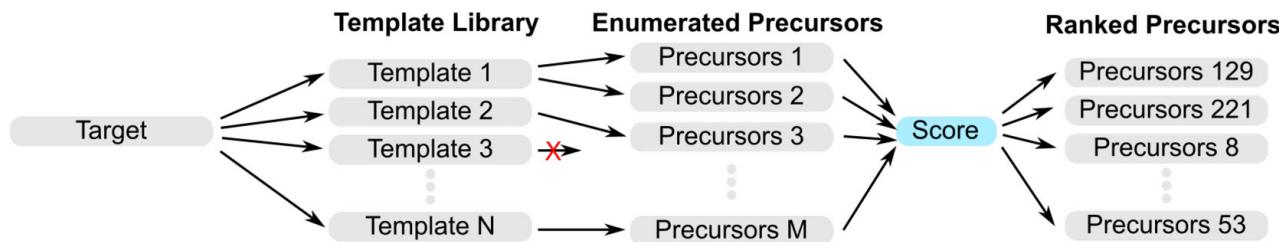
Similarities and differences of present algorithms



- Main categories of retrosynthetic strategies:
 - Template-library based
 - Template free
 - Focused template application

Template-based strategies

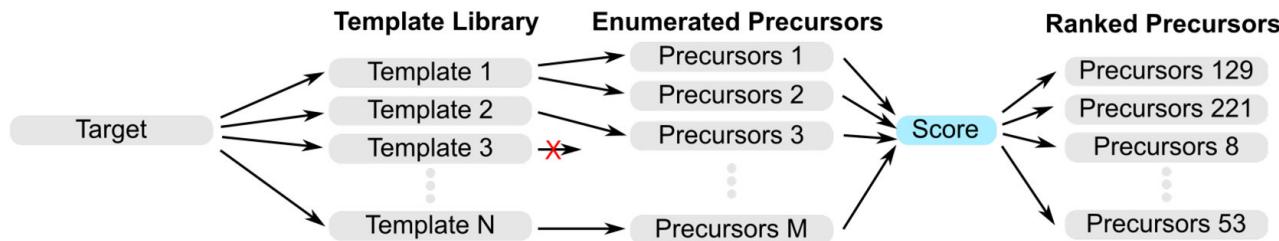
b



- Stored/extracted reaction rules within the algorithm:
 - Chematica: manual input to cover known chemistry (similar to LHASA)
 - Algorithmic template extraction from reaction databases

Template-based strategies

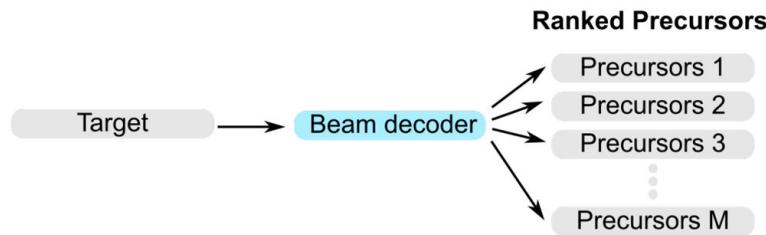
b



- Stored/extracted reaction rules within the algorithm:
 - Chematica: manual input to cover known chemistry (similar to LHASA)
 - Algorithmic template extraction from reaction databases
- Optimization of the pathway generation:
 - Focus only on promising/simplifying disconnections
 - Score/metric to quantify molecular (synthetic) complexity:
 - SMILES length; Chemical Scoring Functions
 - Higher order machine learning approaches (trained with reaction database)

Template free strategies

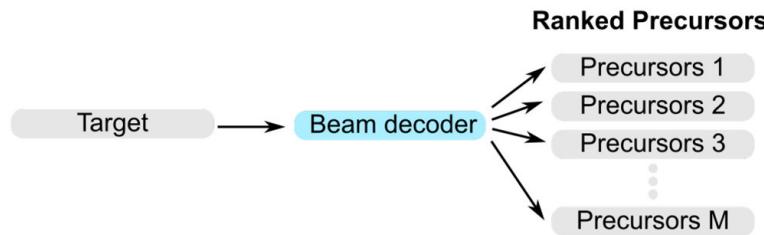
C



- Reaction: sequence prediction using SMILES representations
- "Natural language processing problem": translation of product to reactant

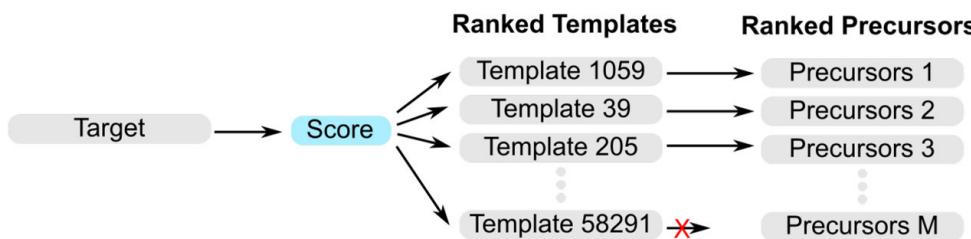
Template free strategies

C



- Reaction: sequence prediction using SMILES representations
- "Natural language processing problem": translation of product to reactant

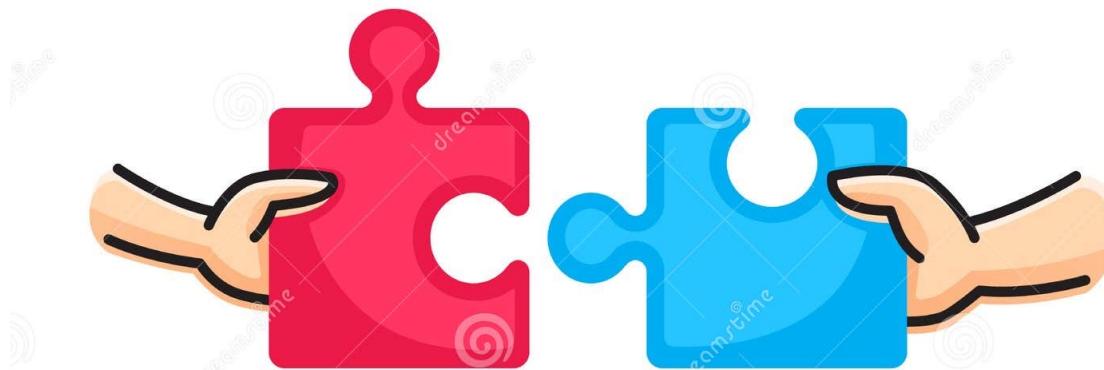
d



- Comparable to how we search for similar compounds on Reaxys
- Ranking template relevance to increase speed
- Structural similarities (Tanimoto distances) to known products are considered

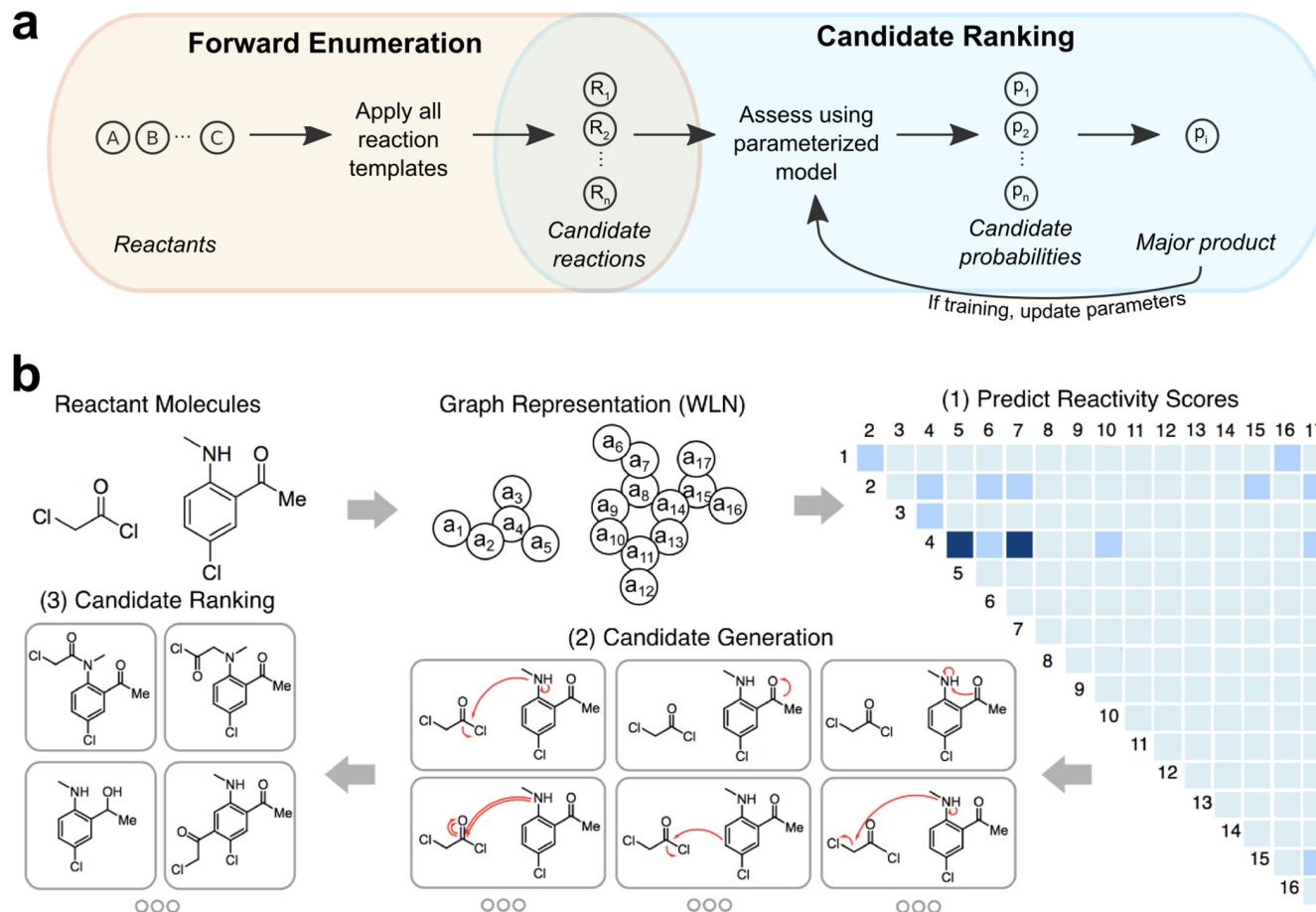
Forward direction: reaction outcome prediction

- Need to be able to evaluate the feasibility of the key steps



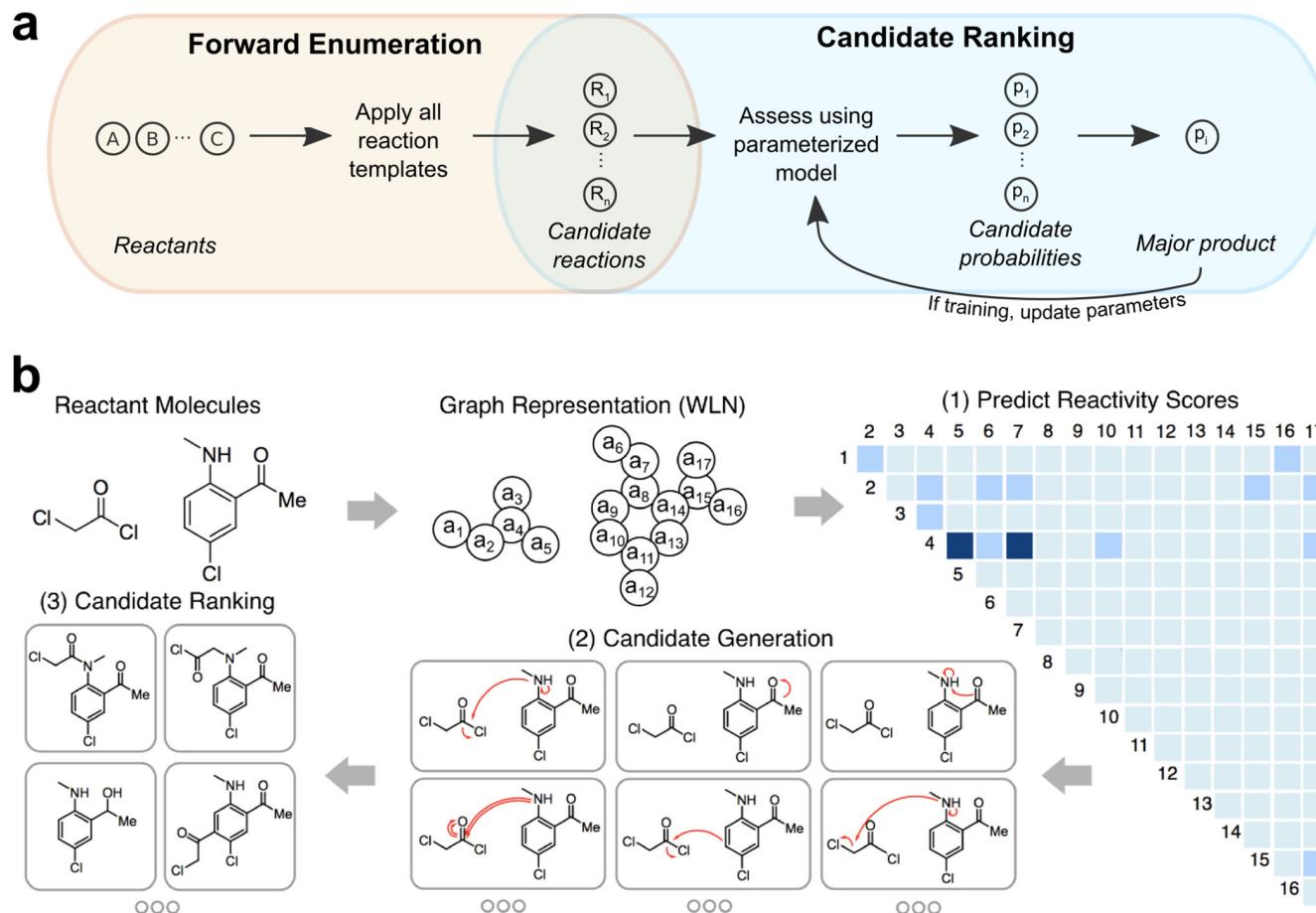
Forward direction: reaction outcome prediction

- Templated-based enumeration of all possible products
- Learn pairwise atom interactions; “microscopic”



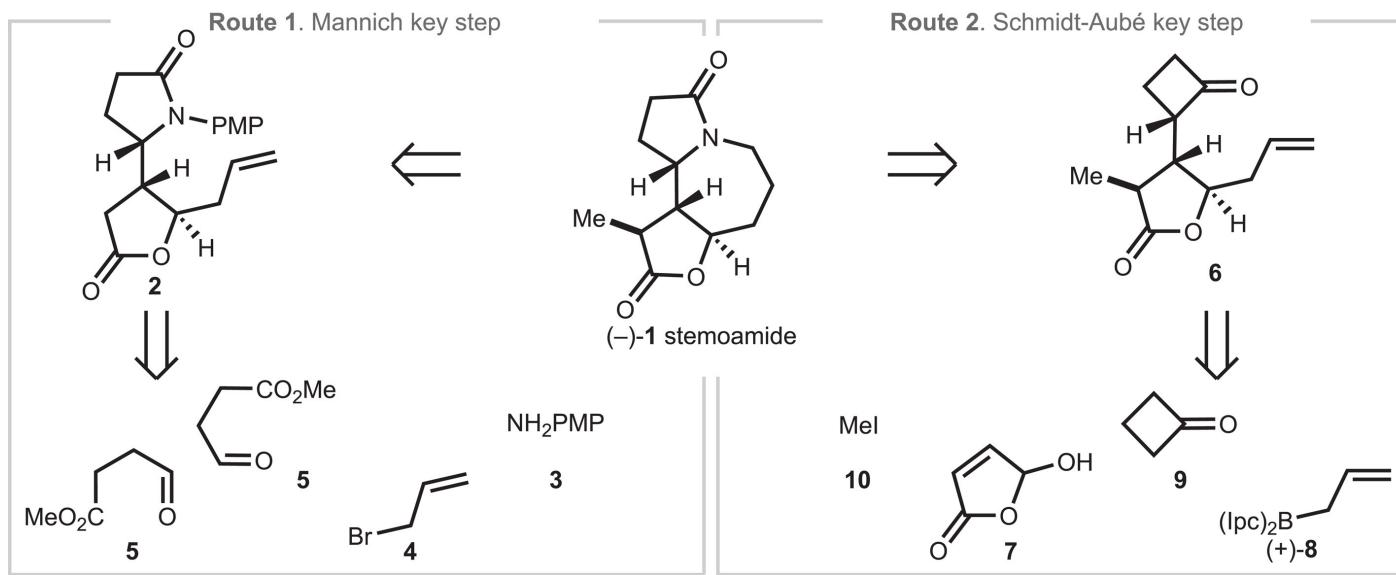
Forward direction: reaction outcome prediction

- Templated-based enumeration of all possible products
- Learn pairwise atom interactions; “microscopic”

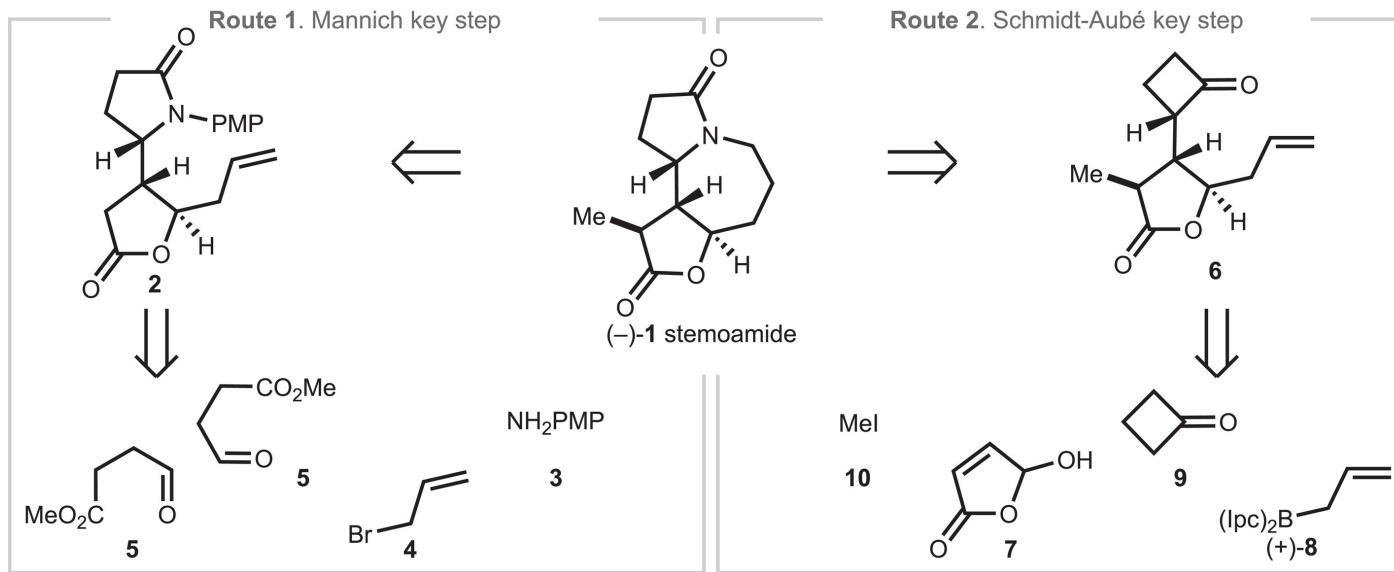


- Yield prediction can be an essential part too!

Case Studies I: Cernak's synthesis of stemoamide



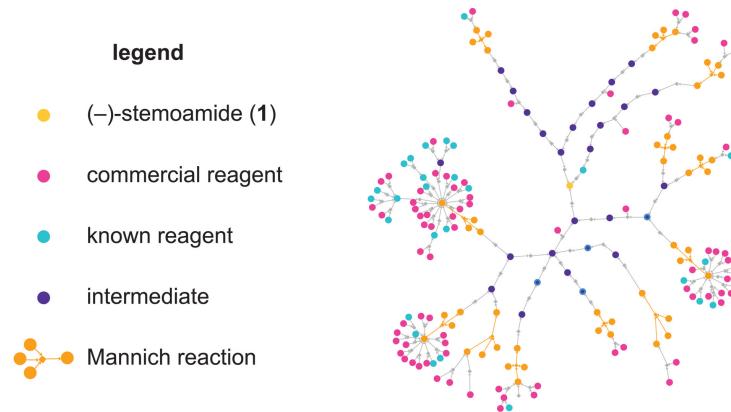
Case Studies I: Cernak's synthesis of stemoamide



- Software used: SYNTHIA
- Mostly using software for the most simplifying key step generation

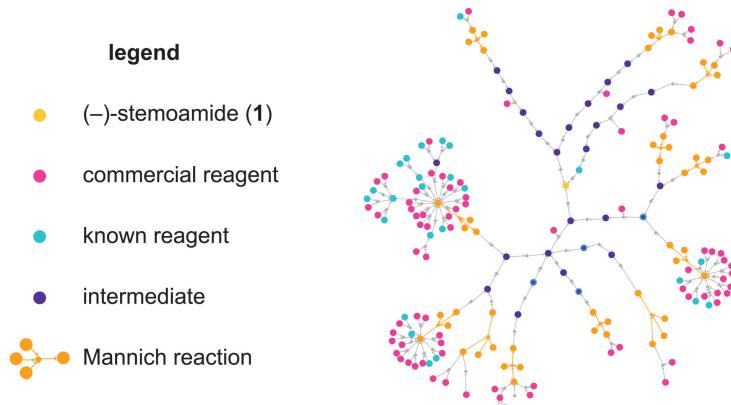
Case Studies I: Cernak's synthesis of stemoamide

- Mannich featured as a consistent step in a lot of proposed routes

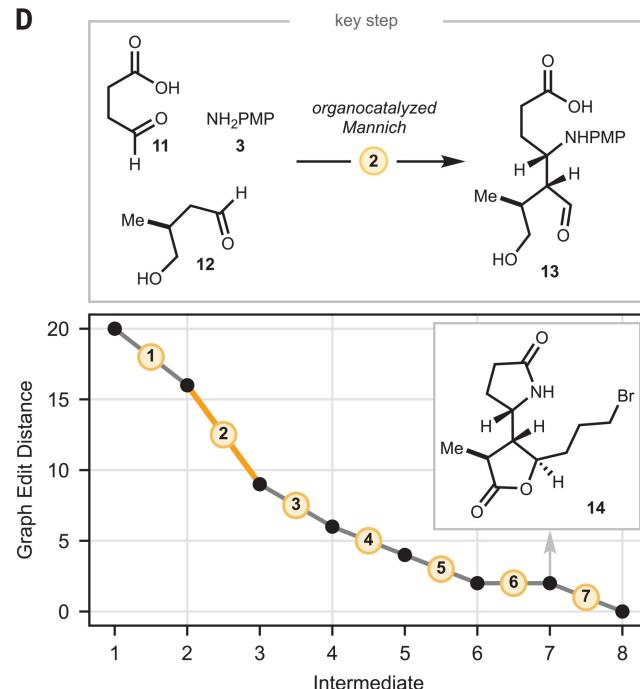


Case Studies I: Cernak's synthesis of stemoamide

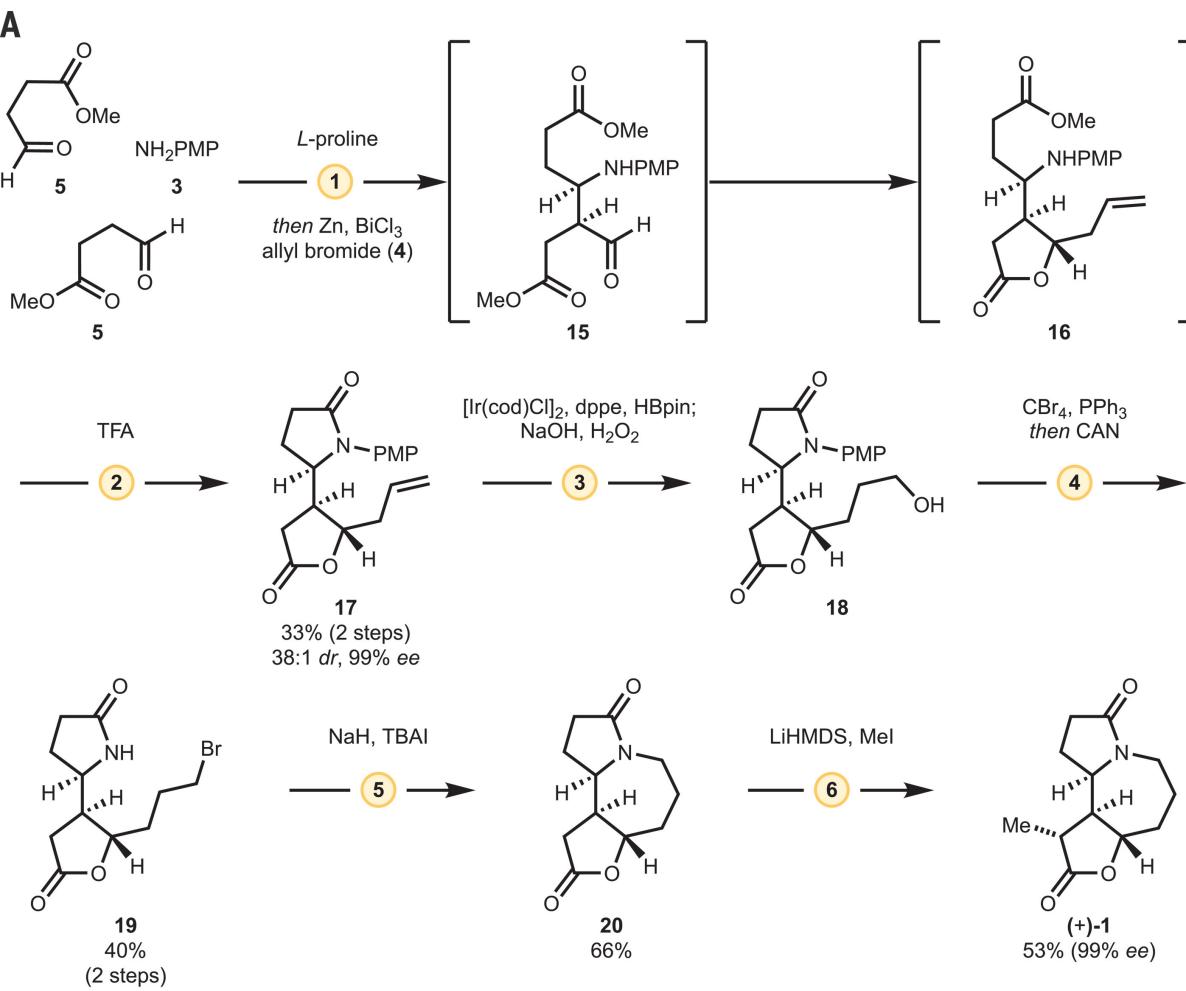
- Mannich featured as a consistent step in a lot of proposed routes



- Mannich also appears to be a highly simplifying step by their scoring analysis



Case Studies I: Cernak's synthesis of stemoamide



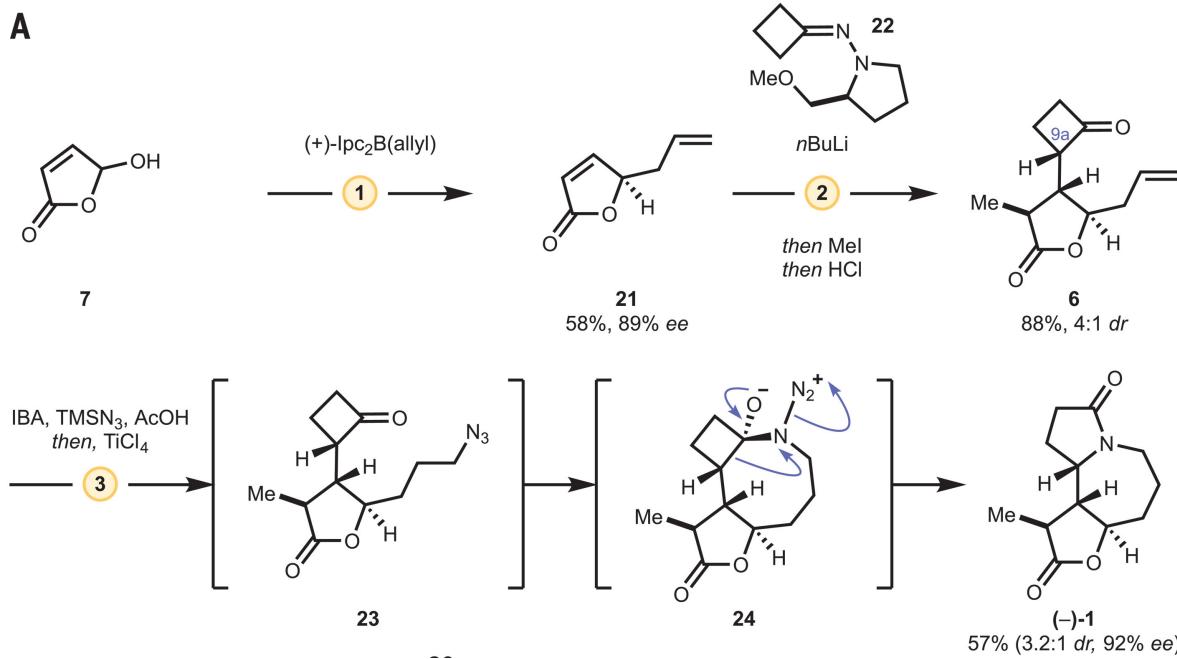
- Changed Mannich step: two equivalents of a more available SM
- Changed sequence of hydrobromination
- Used reported conditions for the endgame: software suggests installing Me early

Case Studies I: Cernak's synthesis of stemoamide

- Repurposing key step: additional results generated; searched using late-stage intermediates

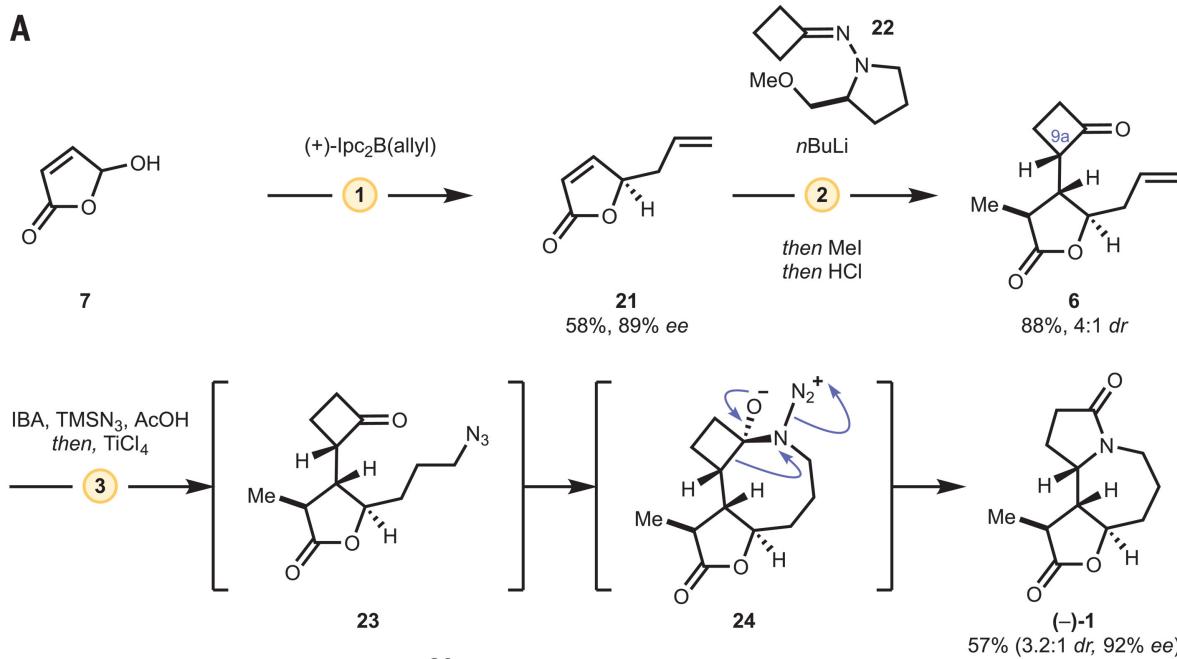
Case Studies I: Cernak's synthesis of stemoamide

- Repurposing key step: additional results generated; searched using late-stage intermediates



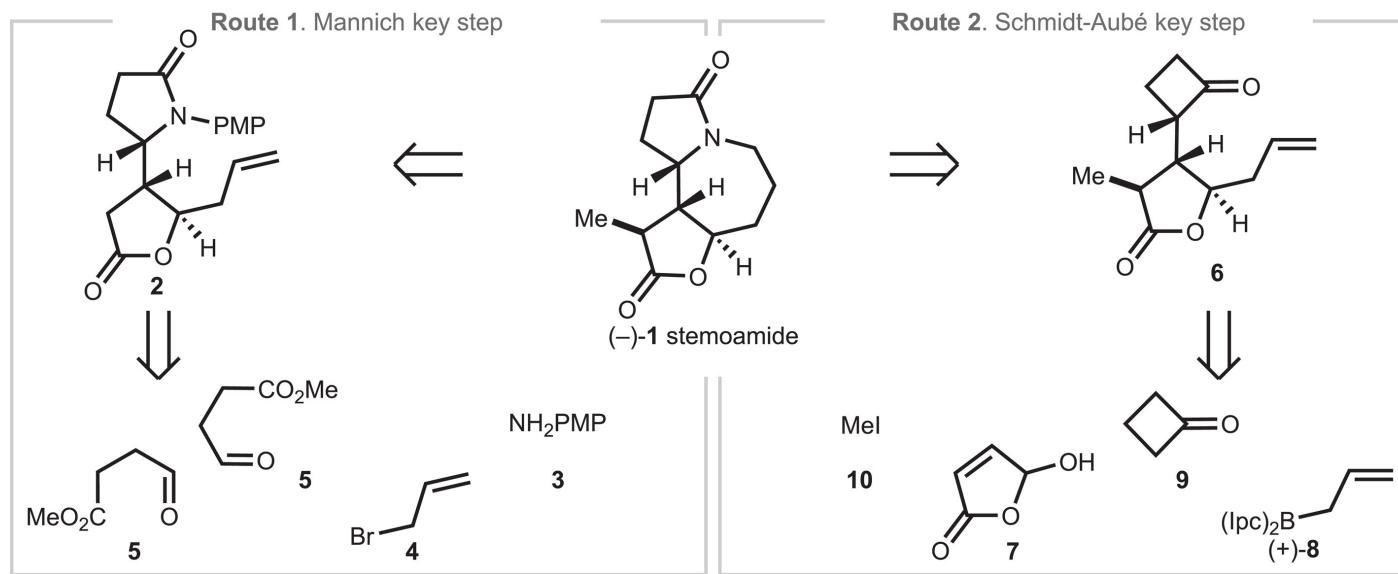
Case Studies I: Cernak's synthesis of stemoamide

- Repurposing key step: additional results generated; searched using late-stage intermediates



- Search performed excluding Mannich steps
- Modified a proposed cyclobutanone intermediate to truncate steps
- SAMP auxiliary inspired by other proposed routes

Case Studies I: Cernak's synthesis of stemoamide

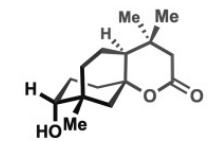


Main Takeaway:

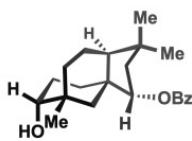
- Retrosynthetic software highly potent in identifying simplifying disconnections
- A chemist's eye can still spot rooms for improvement in intermediate prep
- Graph edit distance analysis—a new way of evaluating routes

Case Studies II: Newhouse synthesis of clovan-2,9-dione

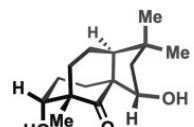
A



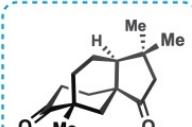
rumphellclovane B (3)



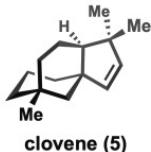
canangaterpenes II (2)
(revised structure)



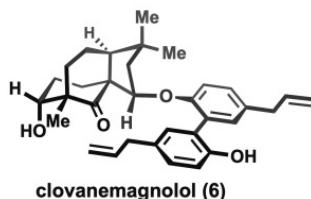
rumphellclovane E (4)



clovane-2,9-dione (1)



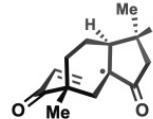
clovane (5)



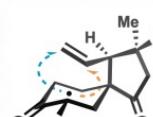
clovane-magnolol (6)

Key radical cyclization evaluated computationally
Baldwin's rules do not inform the feasibility
6-endo or 5-exo ?

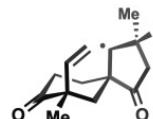
Potential disconnections



7



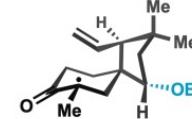
8



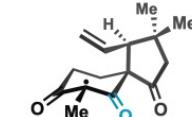
9

Optimal C-C disconnection?

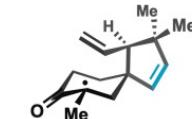
Structural modifications



10



11

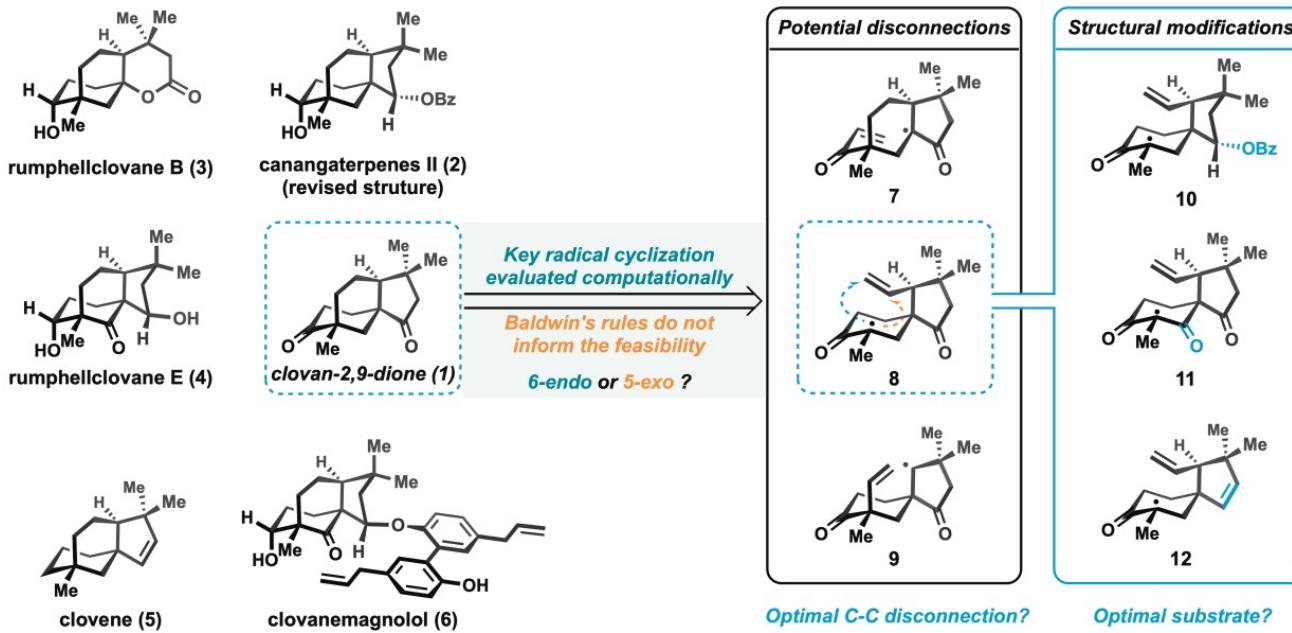


12

Optimal substrate?

Case Studies II: Newhouse synthesis of clovan-2,9-dione

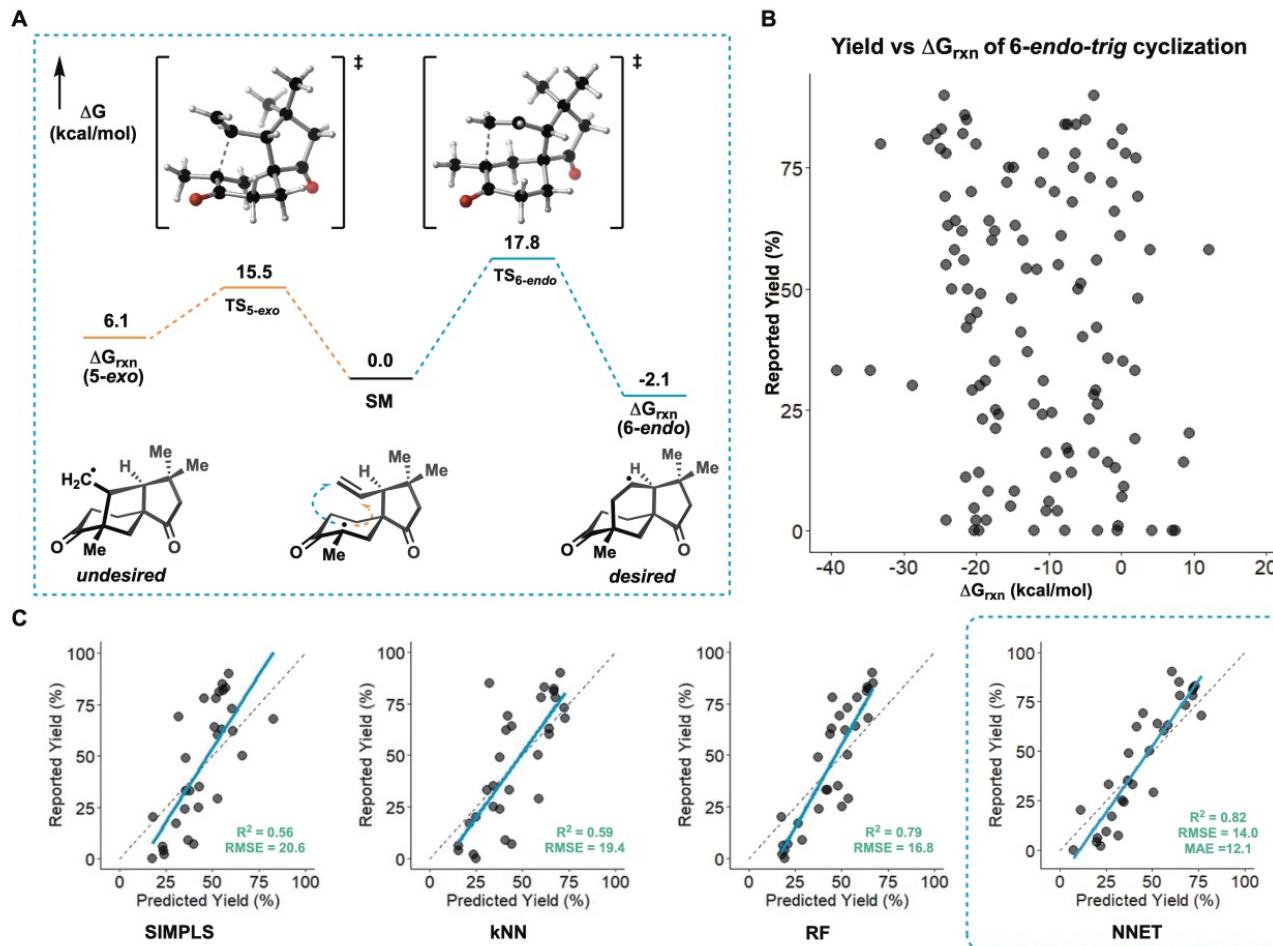
A



- Computational evaluation of the optimal key step (cyclization)
- Machine learning models to find best set of substrates

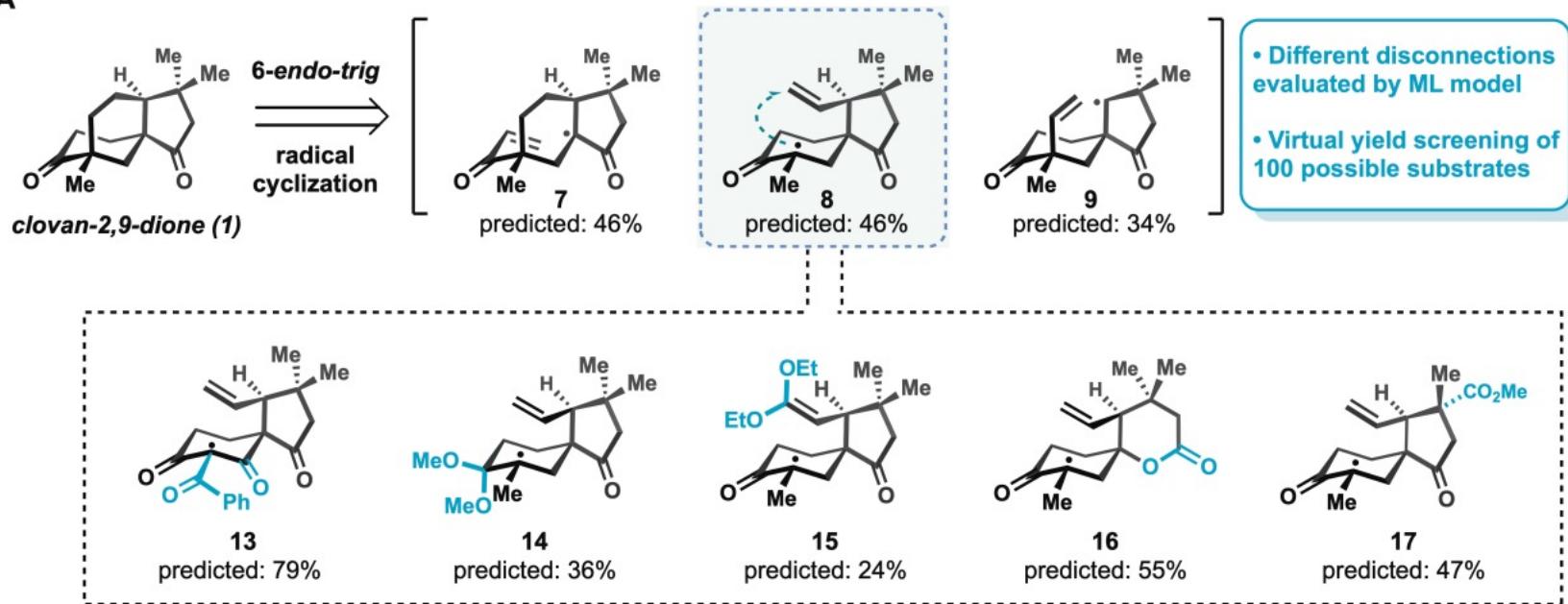
Case Studies II: Newhouse synthesis of clovan-2,9-dione

- Prediction of radical cyclization yield: complicated interplay of kinetics and thermodynamics
- Purely DFT not optimal; ML with literature yield for 6-endo-trig instead



Case Studies II: Newhouse synthesis of clovan-2,9-dione

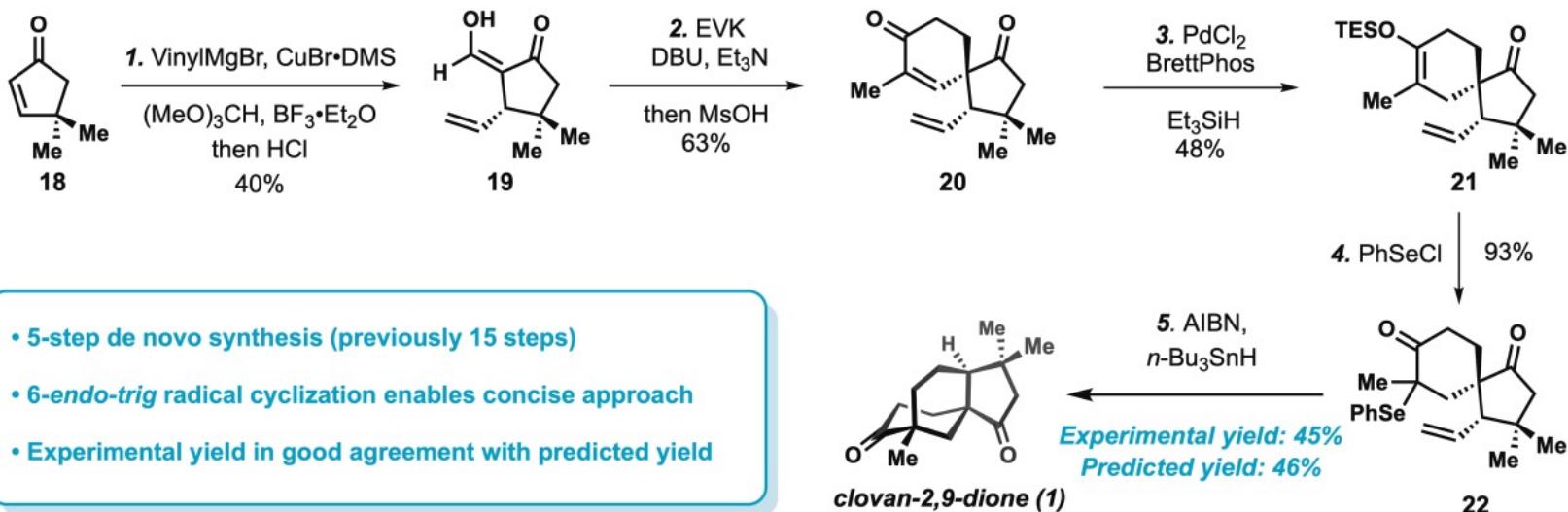
A



- Evaluation of different bond disconnections
- Optimizations by editing the proximal and remote FG's

Case Studies II: Newhouse synthesis of clovan-2,9-dione

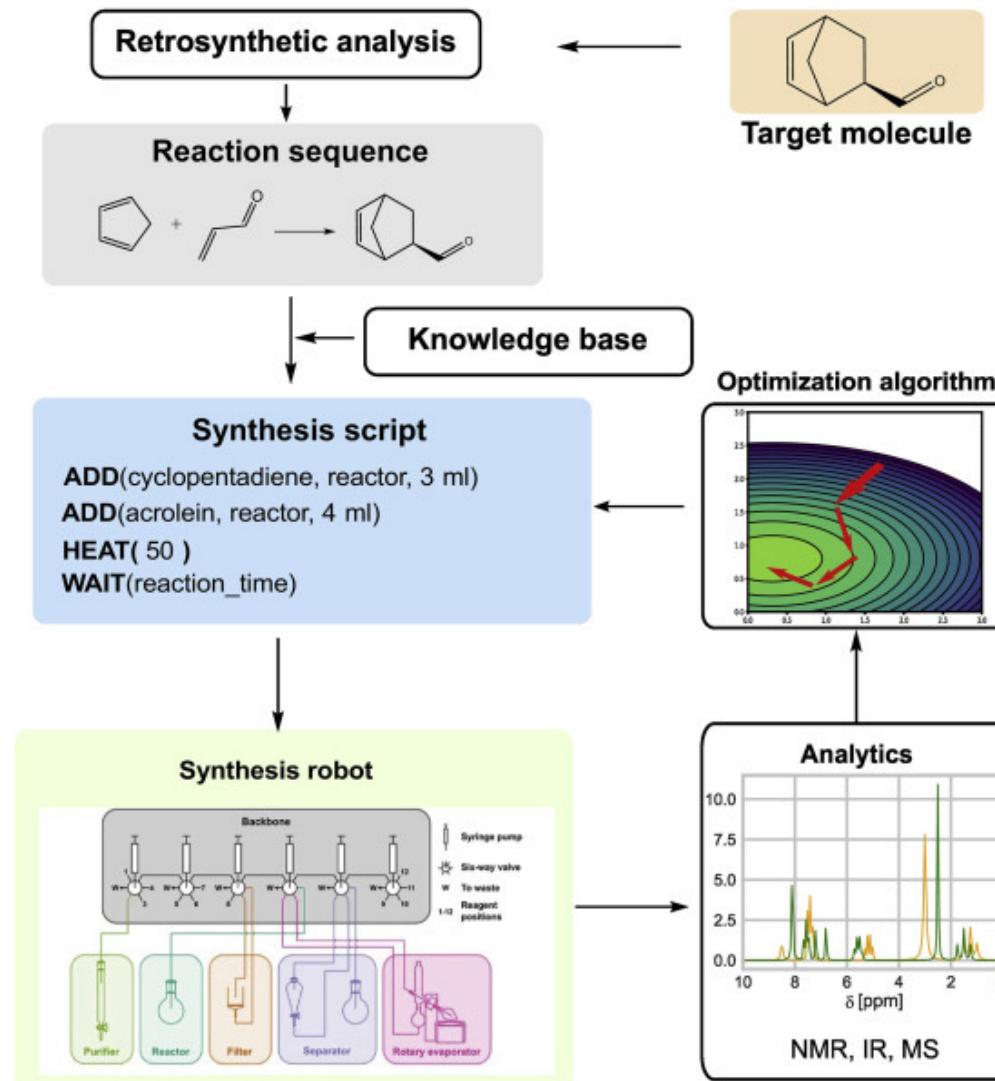
B



- Forward reactivity prediction can also be useful
- Key step evaluation can work in conjunction with route planning

Forward looking: Fully automated chemical synthesis

- The Chemputer: An effort towards a Universal Chemical Synthesizer



Conclusion

