

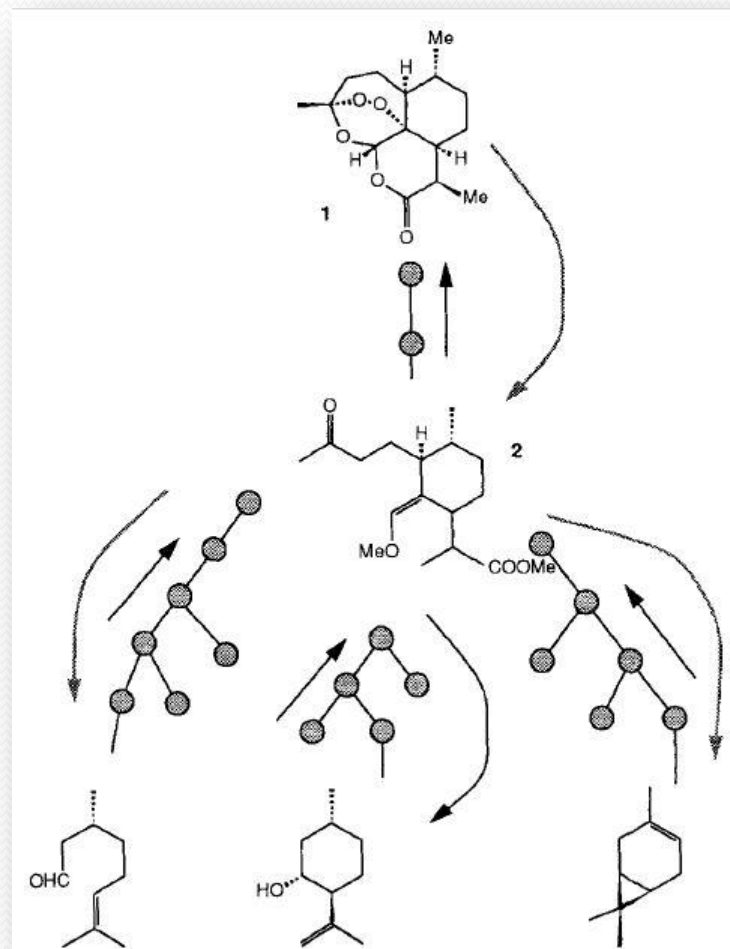
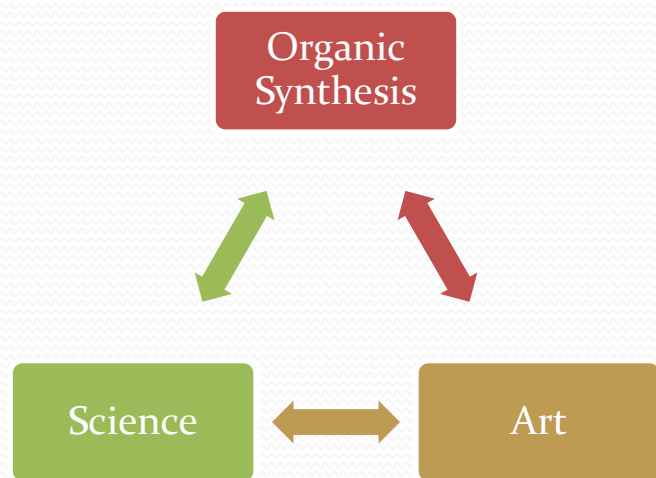
# Computer-Aided Synthesis Design – An Introduction to ARChem

Orr Ravitz & Aniko Simon

June 2011

# Synthesis Design - The Thought Process

- A non linear process
- Driven by knowledge and intuition
- Often biased toward the chemist's specialties and experience
- Knowledge gaps are filled using literature searches



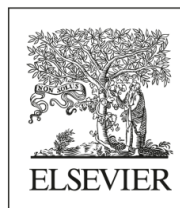
Ihlenfeldt & Gateiger, *Angew.* 1995

# Digital Chemical Data

- Since the mid 1980's chemical data has been extensively digitized
- Major reaction databases are available in electronic format
- Data mining tools offer increasing flexibility in data retrieval, but are not abstracting and extrapolating information



DiscoveryGate®



# Why CASD?

## Chemist

Creativity

Intuition

Strategic perspective

Knowledge (what works,  
and more importantly,  
what doesn't)

## Computer

Thoroughness

Lack of bias

Speed

Low cost

# ARChem

## The Vision:

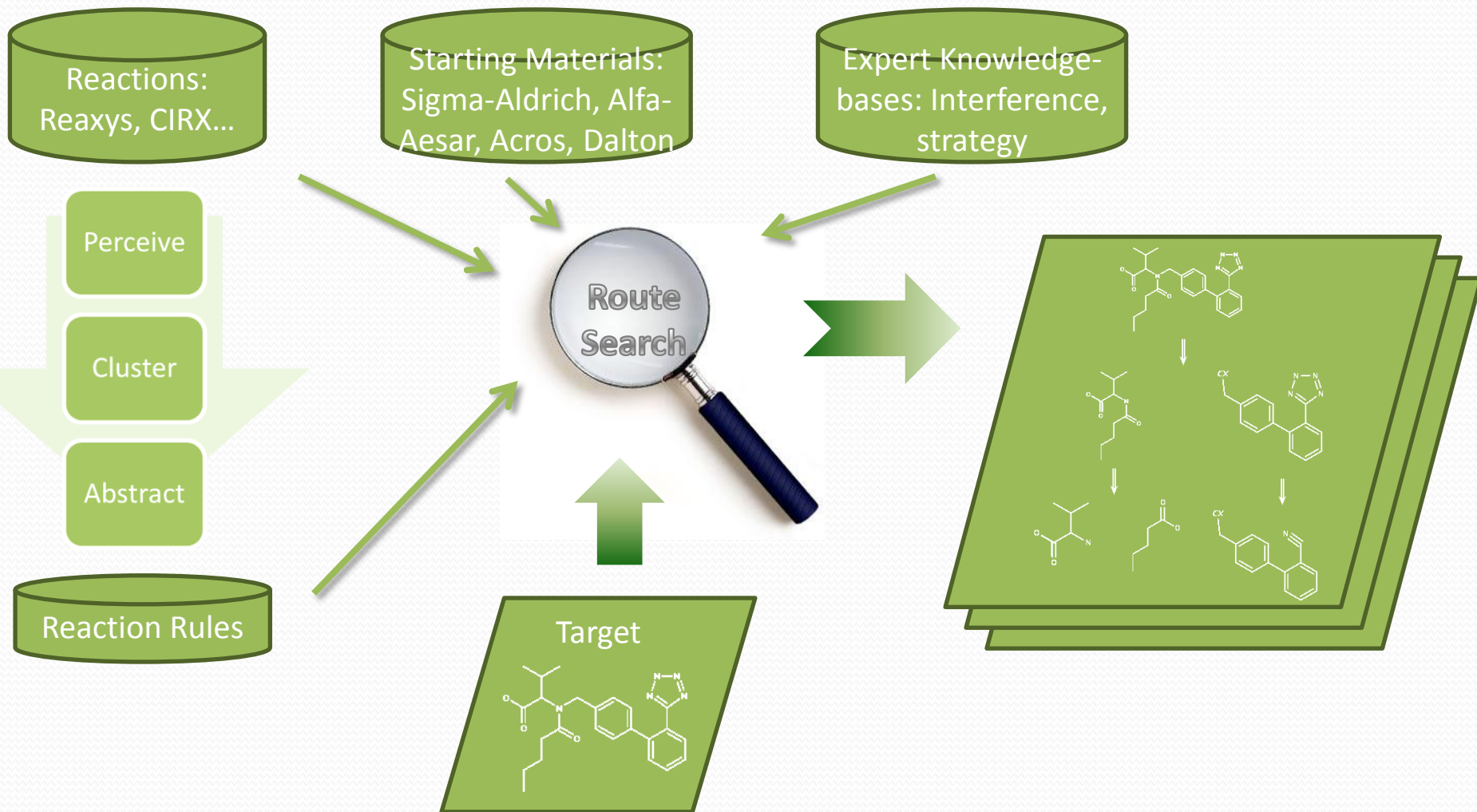
- **Ideas** – Spur new ideas, support creativity
- **Opportunities** – cover greater chemical space
- **Efficiency** – human, synthetic
- **Knowledge** – dissemination and integration

## The Approach:

- Comprehensive rule- and precedent-based retrosynthetic analysis back to available starting materials.
- Automated rule generation with manual rule curation.
- Generate many alternatives.
- Provide supporting literature examples.
- Allow user guidance and control.

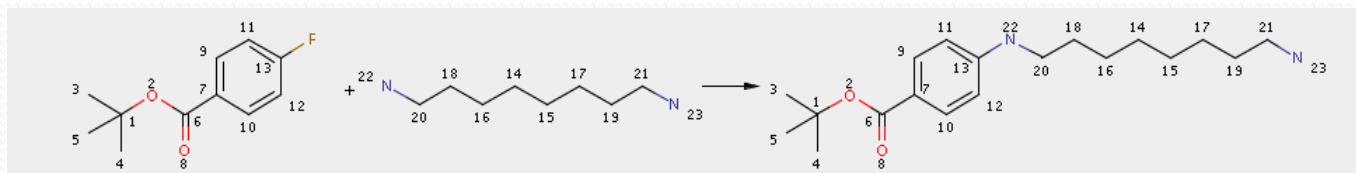


# System Design



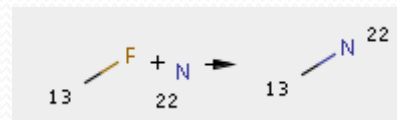
# Reaction Perception

## Source reaction:



Reaction file with atom mapping

## Extracted core



Atoms attached to bonds changed, made or broken in the reaction

## Extended core



Include all structural motifs that are essential for the reaction to occur

Reactions:  
Reaxys, CIRX...

Perceive

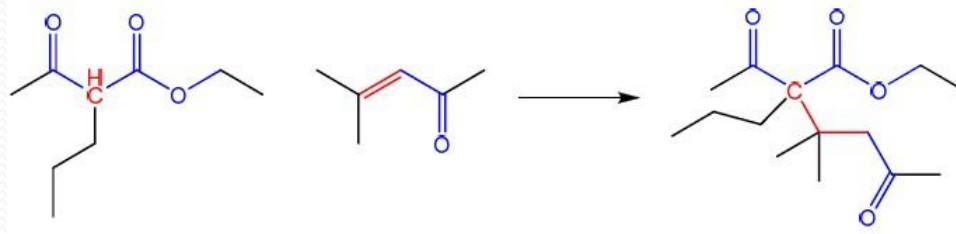
Cluster

Abstract

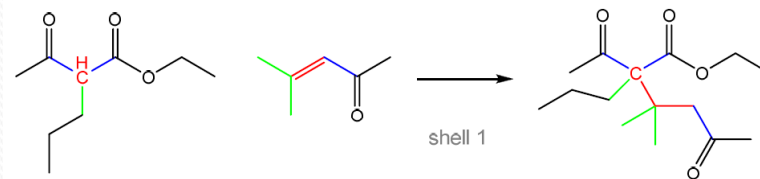
Reaction Rules

# Extending the Core: Passengers vs Drivers

The goal of chemical perception is to discriminate between structural features that are essential for the reaction, and those that are passengers.

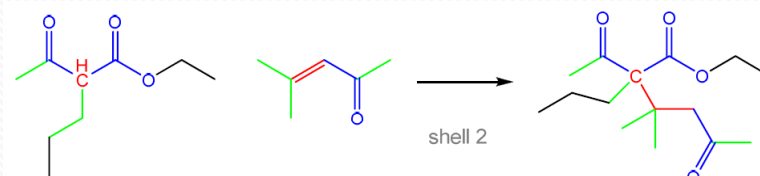


Shell-based approach: 1<sup>st</sup> shell



Too generic

2<sup>nd</sup> shell



Too specific

Graph-based methods are inappropriate.

# Rule Extraction

## Source reactions

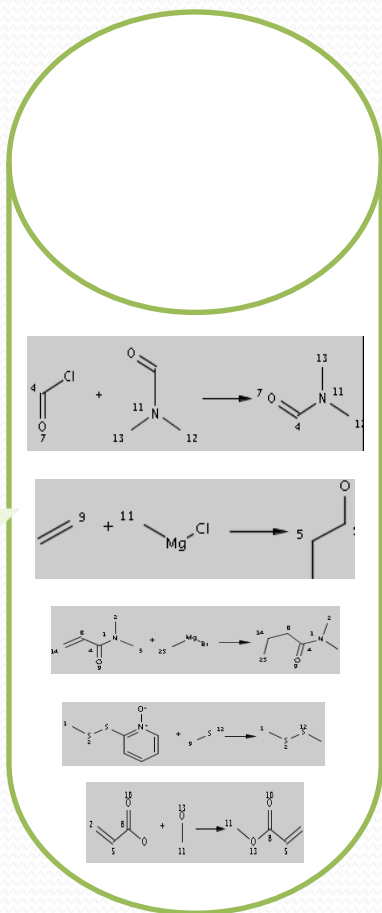
Reactions:  
Reaxys, CIRX...

Perceive

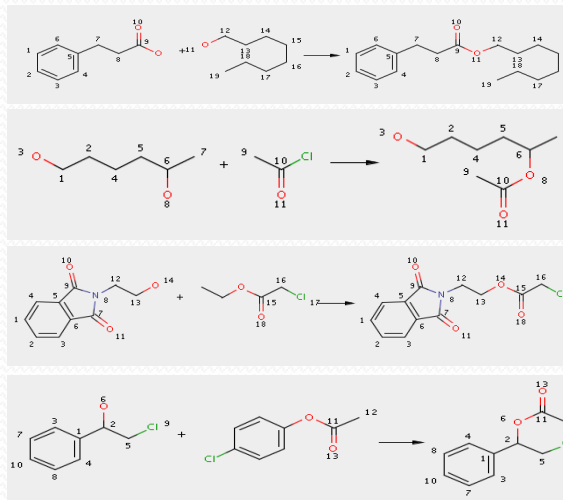
Cluster

Abstract

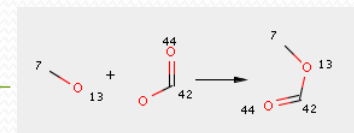
Reaction Rules



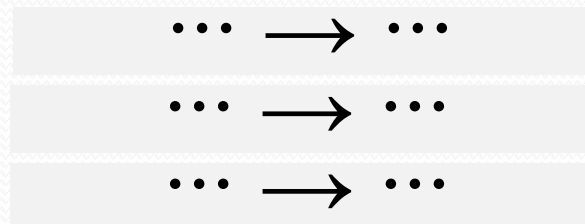
## Esterification examples



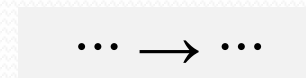
## Esterification rule



## Other examples



## Other rule



# Rule Extraction

Reactions:  
Reaxys, CIRX...

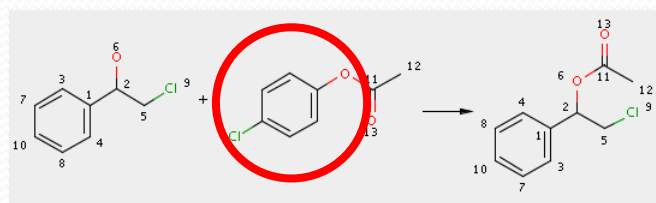
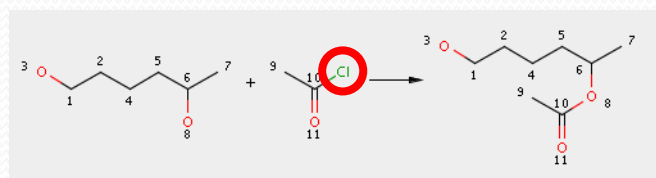
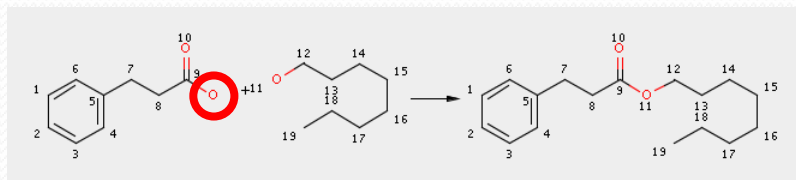
Perceive

Cluster

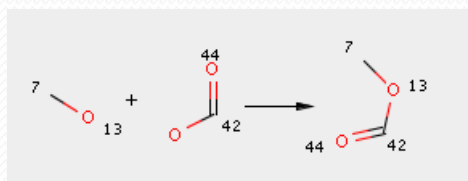
Abstract

Reaction Rules

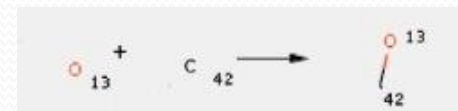
## Similar extended cores



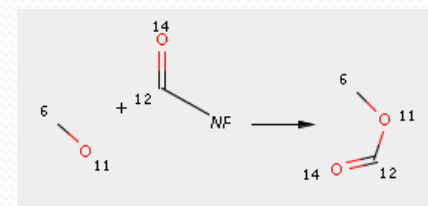
## Completed reaction rule



## Common extracted core



## Generalized rule



Nucleofuge (NF) -  
a leaving group which  
carries away the bonding  
electron pair.

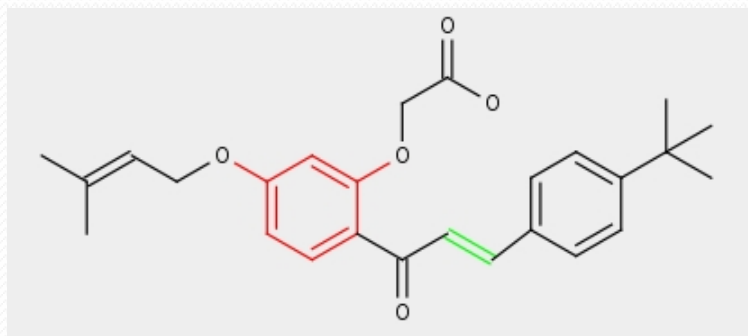
Generalized group (NF) is  
replaced by the most  
common group.

# The Search

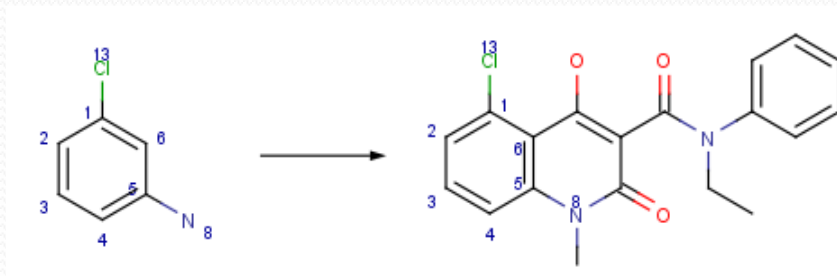
- The search is exhaustive:
  - All the source reactions are used – “Exact Match”.
  - Rules are invoked according to an “example count” user-defined parameter. Only rules based on more examples than the value of the parameter are used in search.
- Means to control combinatorial explosion:
  - Rule categorization and category-based prioritization of rule application.
  - Synthetic depth.
  - Example count – determines the number of rules to be used.
  - User guidance – directing and limiting the search.

# User Guidance

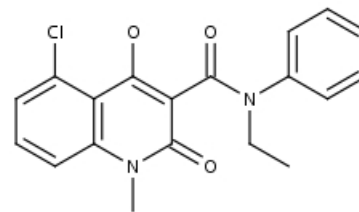
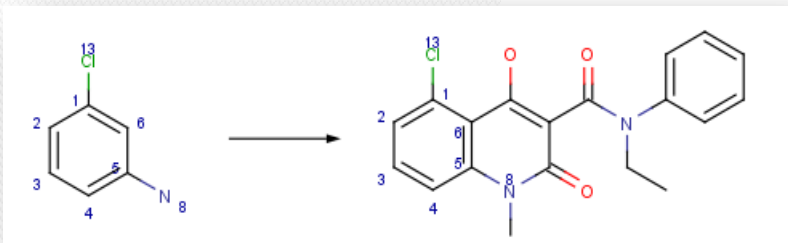
Preserving and targeting bonds



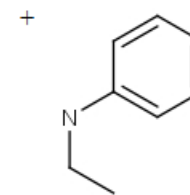
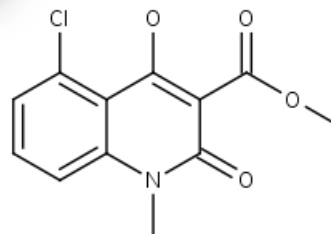
Starting material oriented search



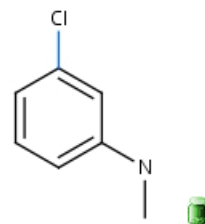
Control the combinatorial explosion and guide the search



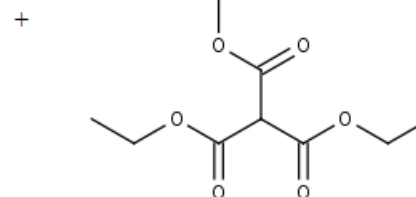
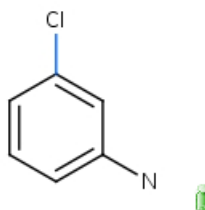
✓ 98% ↓ 1 of 46



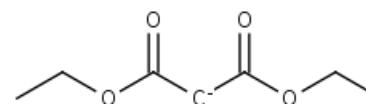
12 examples, 53% ↓ 20 alternatives



✓ 10344 examples, 60% ↓ 26 alternatives



✓ 90% ↓ 7 alternatives



[User Starting Material]

# Presenting the results

## The main challenges:

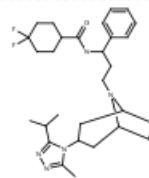
- The solutions space could be large.
- Some tension between viewing full routes and individual steps.
- Enormous amount of information in various formats.

## Adopted principles:

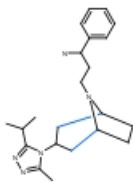
- Viewing is structure-driven (little text).
- Hybrid step-by-step / full route display – route construction via ‘Solutions Walking’.
- Supplemental information available by mouse-click.
- Solutions space pruning – elimination of compounds and transforms.
- Modifiable fields – yield, price, starting materials.



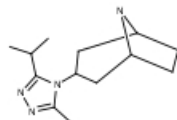
The chemist is in control!



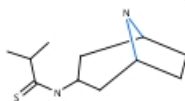
✓ 66075 [examples](#), 49% ↓ 40 alternatives



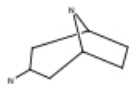
5019 [examples](#), 64% ↓ 32 alternatives



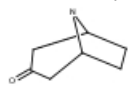
269 [examples](#), 61% ↓ 15 alternatives



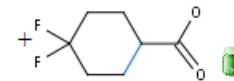
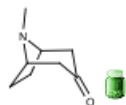
805 [examples](#), 73% ↓ 6 alternatives



750 [examples](#), 63% ↓ 4 alternatives

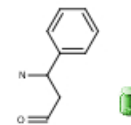


✓ 100% ↓ 4 alternatives

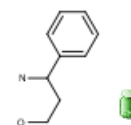


↓ 12 alternatives

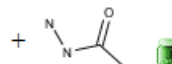
+



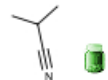
✓ 22548 [examples](#), 80% ↓ 34 alternatives



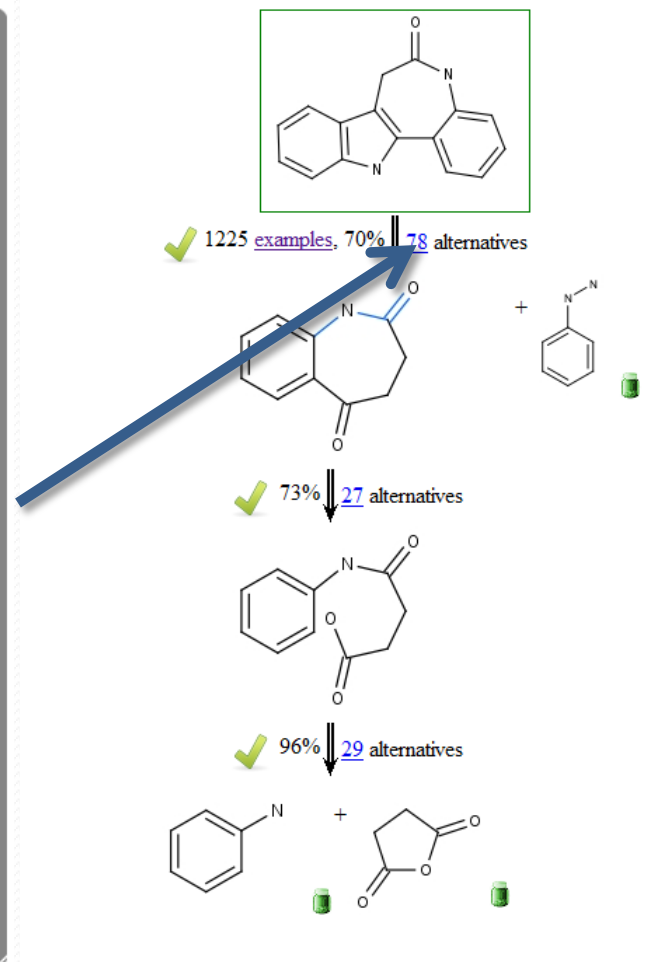
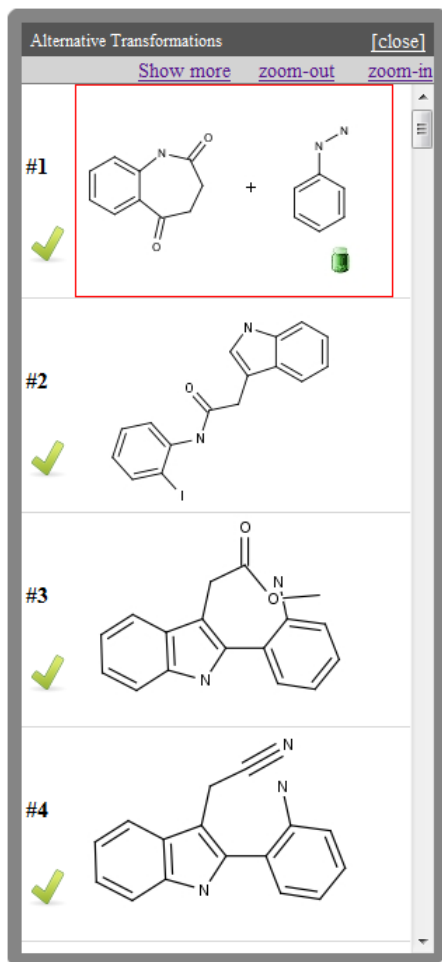
↓ 28 alternatives



✓ 90% ↓ 11 alternatives



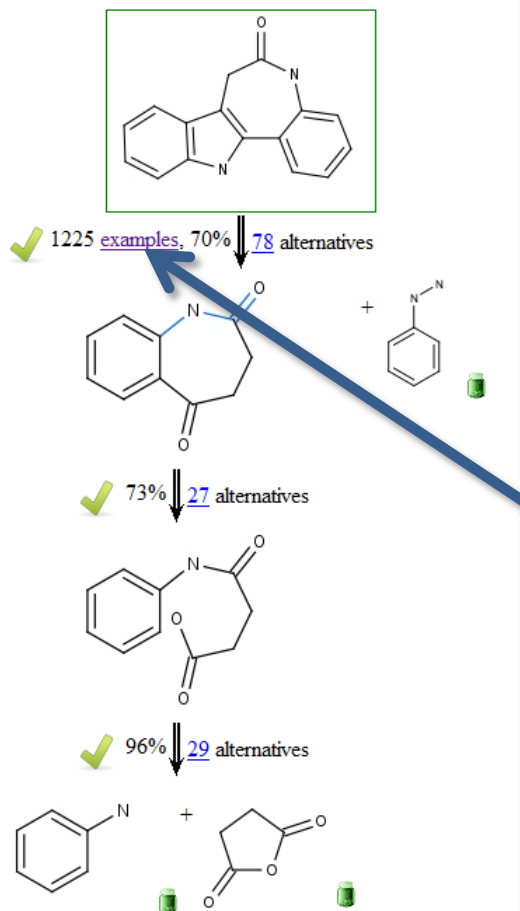
# Viewing the Solutions



- Alternative transforms ordered by confidence.
- Hovering over an alternative shows the top rank subtree.

# Viewing the Solutions

- Literature examples support suggested transformations.
- Easy access to Reaxys to view full record.



Rules [close]

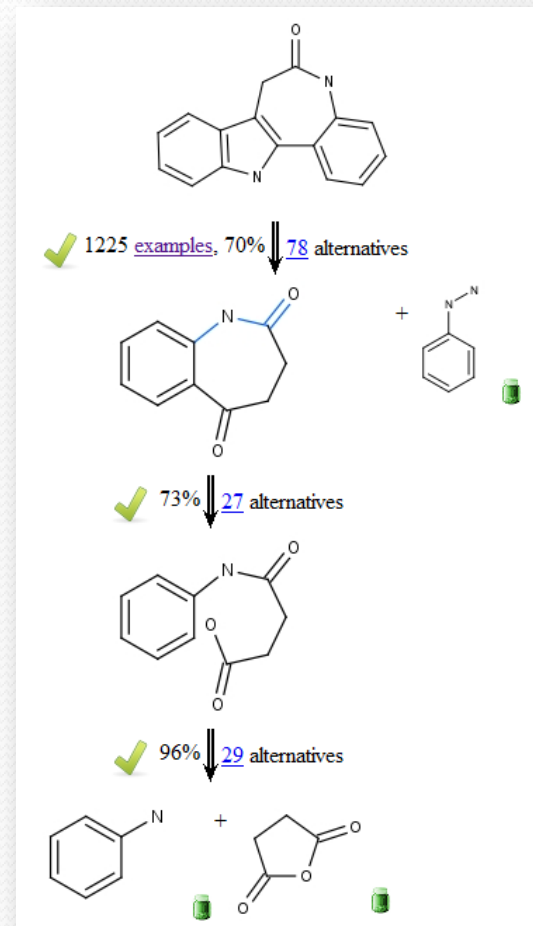
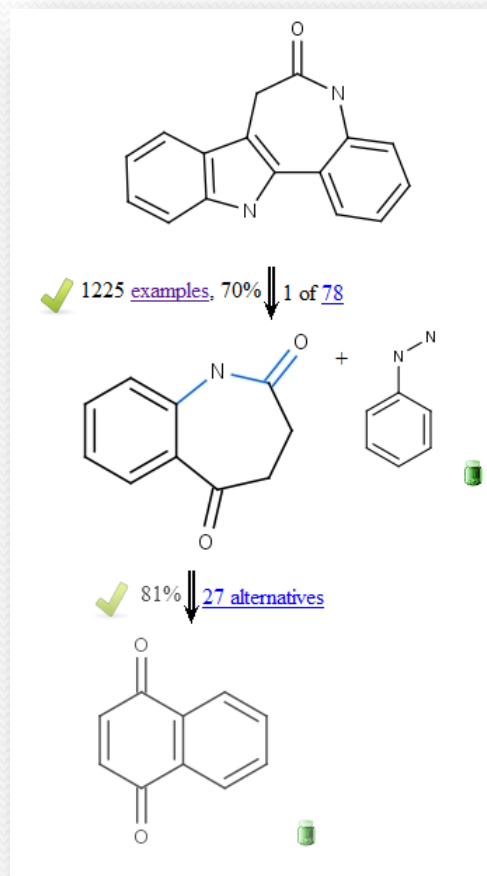
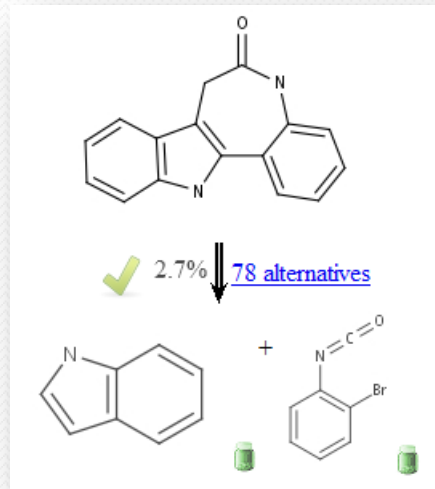
Rule #3: Reaxys[1361408]: Unnamed

1 of 3

Sorted 1 of 984 Export

Reaction details: Reaxys	
ReaxysID	<a href="#">3394905</a>
Conditions	1.) glacial acetic acid, 70 deg C, 1 h, 2.) 70 deg C, 1 h
External ID	3394905
Product	pauellone
Reactant	1H-<1>benzazepine- 2,5(3H,4H)-dione//N- Phenylhydrazine
Reagent	2.) concd. H2SO4
Reaction Name	Preparation
Comments	Yield given. Multistep reaction
Reagent	2.) ZnCl2

# Preferred Synthesis Construction



# The human factor

CASD “has been met with utter scepticism. even hostility, from the majority of chemists” (Ihlenfeldt & Gasteiger 1995)``

Lacking accuracy

Limited scope

Cumbersome design

Misunderstanding of concept

intimidation

## The challenges:

- ▶ Chemists are very knowledgeable 😊
- ▶ Chemists use data retrieval systems with low margin of error.
- ▶ CASD as a predictive tool will err – it is not better than an experienced chemist.
- ▶ CASD systems are new and different – need to be integrated into workflow.

## The way forward:

- ▶ Develop with the users.
- ▶ Identify niches in the synthesis development process in which CASD systems are likely to have the greatest impact.
- ▶ Design the tool to support creativity, not as an alternative.
- ▶ Increase integration with other computational tools.

# Some Conclusions

- CASD remains a great scientific challenge.
- Advances in databases, computer hardware and computational methods set the stage for a comeback of CASD.
- The system offers another avenue to search databases and convert data into knowledge.
- Software should be designed as an assisting tool rather than a black box.
- A collaborative development with users is essential (contact us for details).
- Computers already play a role in the synthesis planning – the end users may be more approving to CASD (?)

Thank you

[www.simbiosys.com](http://www.simbiosys.com)  
[archem@simbiosys.com](mailto:archem@simbiosys.com)