NIBR
NX-SIGMA
Cheminformatics

Novartis Chemical Universe Searching (Astronomically) Large Spaces

Brian Kelley
RDKit UGM
October, 2016

NOVARTIS

What is the NCU?

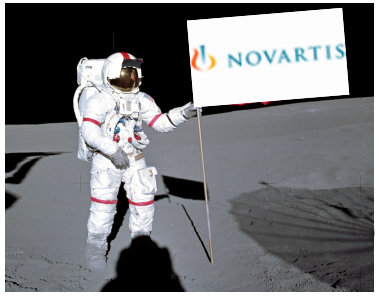
What are we searching for?

New Heaven and Earth

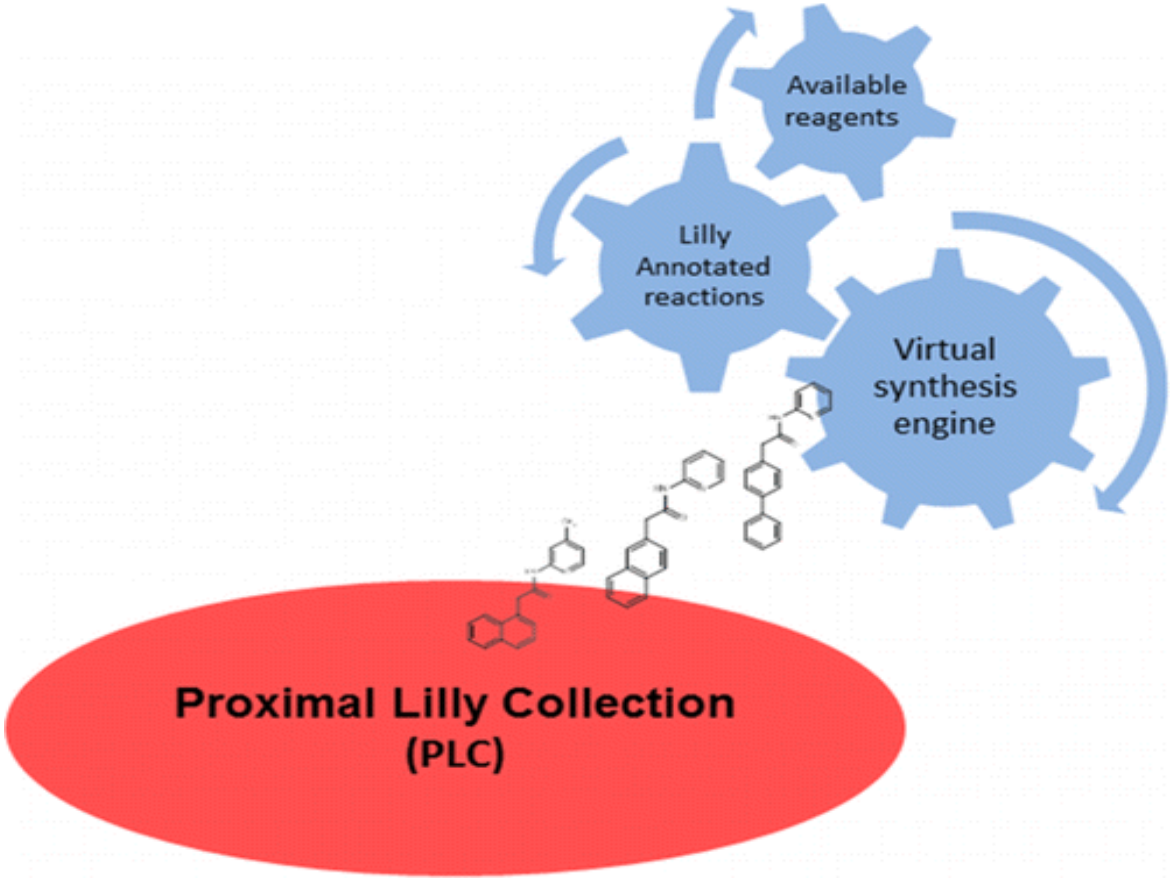
I, new-awakened, with my hand stretching out and touching the unknown, the real unknown, the unknown unknown.

DH Lawrence (not Rumsfeld)

I claim this land for Novartis



We are not alone in the universe



The Proximal Lilly Collection: Mapping, Exploring and Exploiting Feasible Chemical Space.
Nicolau C et al.
J Chem Inf Model. 2016 Jun 27

Known Knowns

Internal/External Screening compounds 10s of millions

DNA Encoded Libraries 100s of millions

Available reagents 100s of thousands

Reagents used sometime-somewhere 1s of millions

Known Knowns

Reactions

- Functional Groups (Aldehyde/Carboxylic acid …)
- 59 Robust Reactions
- (internal) Condensed Multistep reactions (~1K)
- (External) Photochemistry, C-H Activation
- Lilly annotated reaction set

A Collection of Robust Organic Synthesis Reactions for *In Silico* Molecule Design
J. Chem. Inf. Model., **2011**, *51* (12), pp 3093–3098
DOI: 10.1021/ci200379p

Known Unknowns

Is a reagent compatible in a reaction?

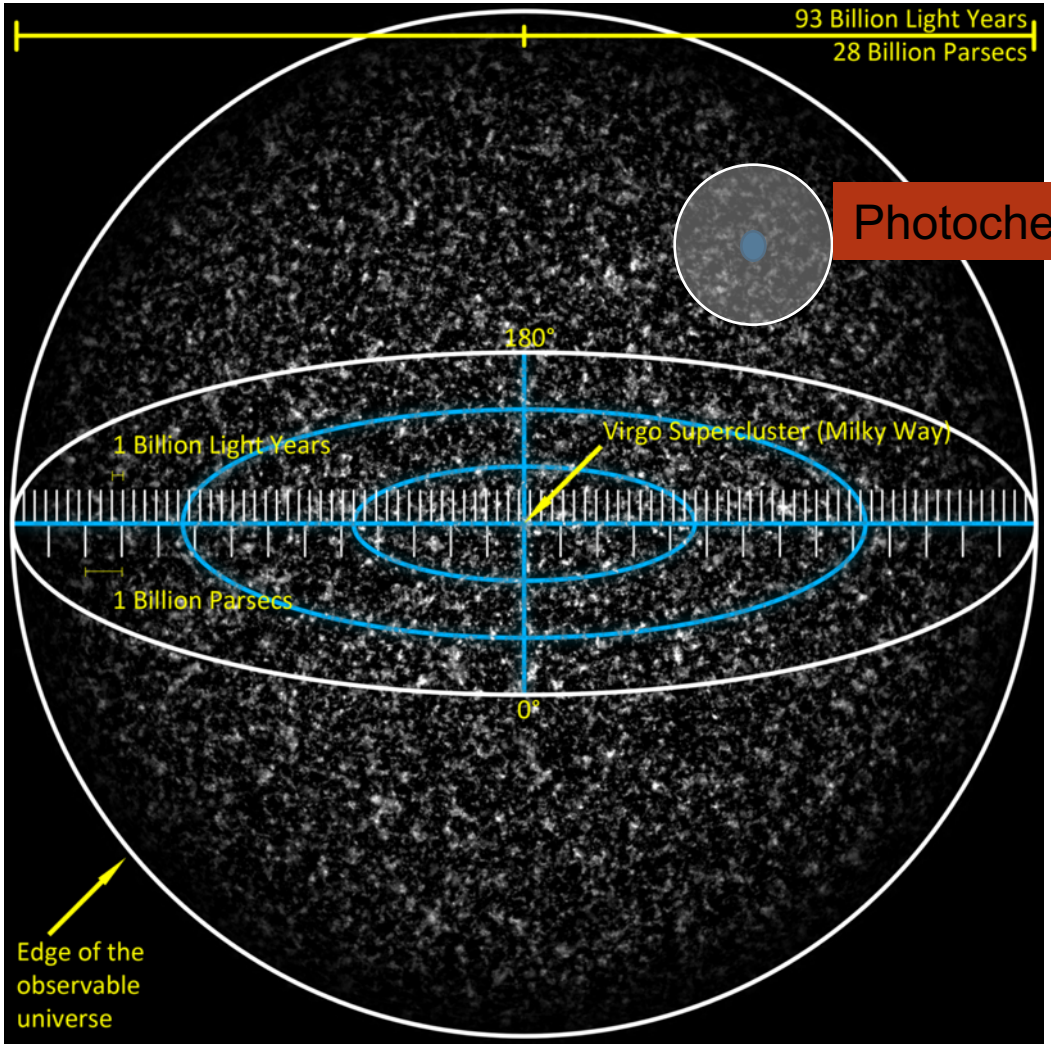
- Can we use new reagents with reactions we have already done?
- If so will the products be useful in products?
  - Soluble?
  - Drug Like?
  - Active?
  - Novel?

Is the NCU feasible?

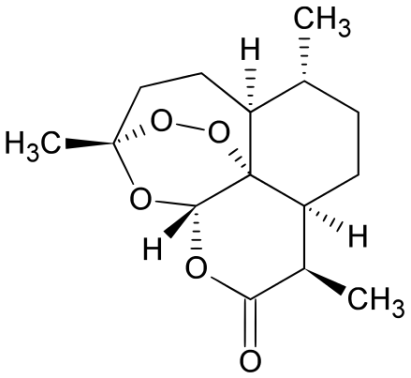
- What is the utility?
- Who enters new reactions?

Unknown Unknowns

Source wikipedia



Photochemistry



NOVARTIS

What is a reaction?

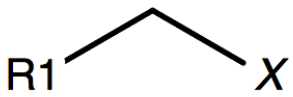
Electronic representation of chemical reality

- Various levels of chemical fidelity
- Most electronically described reactions are only valid within a smallish window of reagents.

What is a reaction?

Negishi Coupling

organohalide

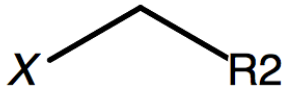


alkenyl, aryl, allyl, alkynyl or propargyl

chloride, bromide, or iodide, (triflate and acetyloxy)

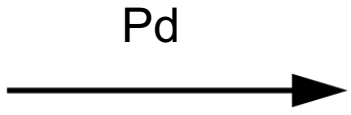
+

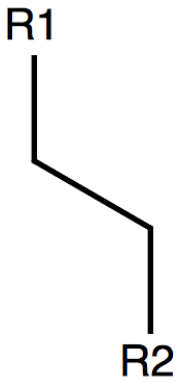
organozinc



Zn + (chloride, bromine or iodine)

alkenyl, aryl, allyl, alkyl benzyl, homoallyl, and homopropargyl.





C-C coupling

What is a reaction?

Negishi Coupling (SMARTS Encoding)

[#6;$([#6]~[#6]);!$([#6]~[S,N,O,P]):1][Cl,Br,I].

[Cl,Br,I][#6;$([#6]~[#6]);!$([#6]~[S,N,O,P]):2]

>>[#6:2][#6:1]

Got it?

What is a reaction?

Negishi Coupling (SMARTS Encoding)

```
[#6;$([#6]~[#6]);!$([#6]~[S,N,O,P]):1][Cl,Br,I].
[Cl,Br,I][#6;$([#6]~[#6]);!$([#6]~[S,N,O,P]):2]
>>[#6:2][#6:1]
```

Reagent1

Reagent2

Product
(C-C coupling)

NOVARTIS

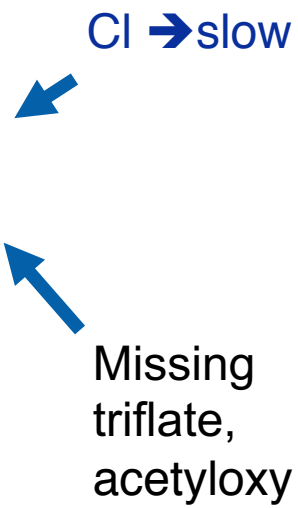
What is a reaction?

Negishi Coupling (SMARTS Encoding)

[#6;$([#6]~[#6]);!$([#6]~[S,N,O,P]):1][Cl,Br,I].

[Cl,Br,I][#6;$([#6]~[#6]);!$([#6]~[S,N,O,P]):2]

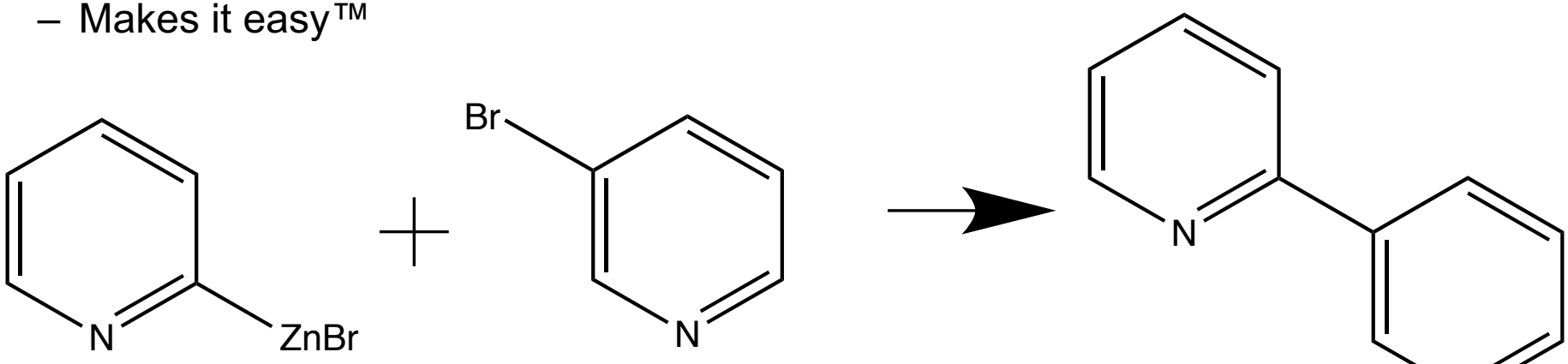
>>[#6:2][#6:1]



What is a reaction?

Most reactions in the NCU are so called template based reactions

- A+B => C
- Makes it easy™



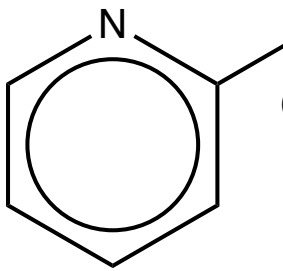
organozinc

organohalide

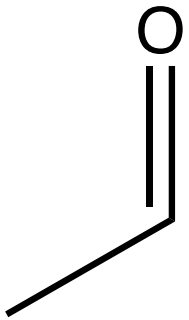
What is a reaction?

Risks

- Enough information to get the known reagents, not enough information to know the incompatible.
- Hydrogens (and electronic environments) are tricky
  - For now, lets say the NCU is an “idea” generator backed by good but not great knowledge of chemistry.



Can R be a hydrogen?



Is that
really
an aldehyde?

Reactions in RDKit Space

MD Reaction files

Smarts based reactions (not smirks, but superset)

New RDKit Reaction tools

SanitizeRXN - simple fixes for common reaction failures

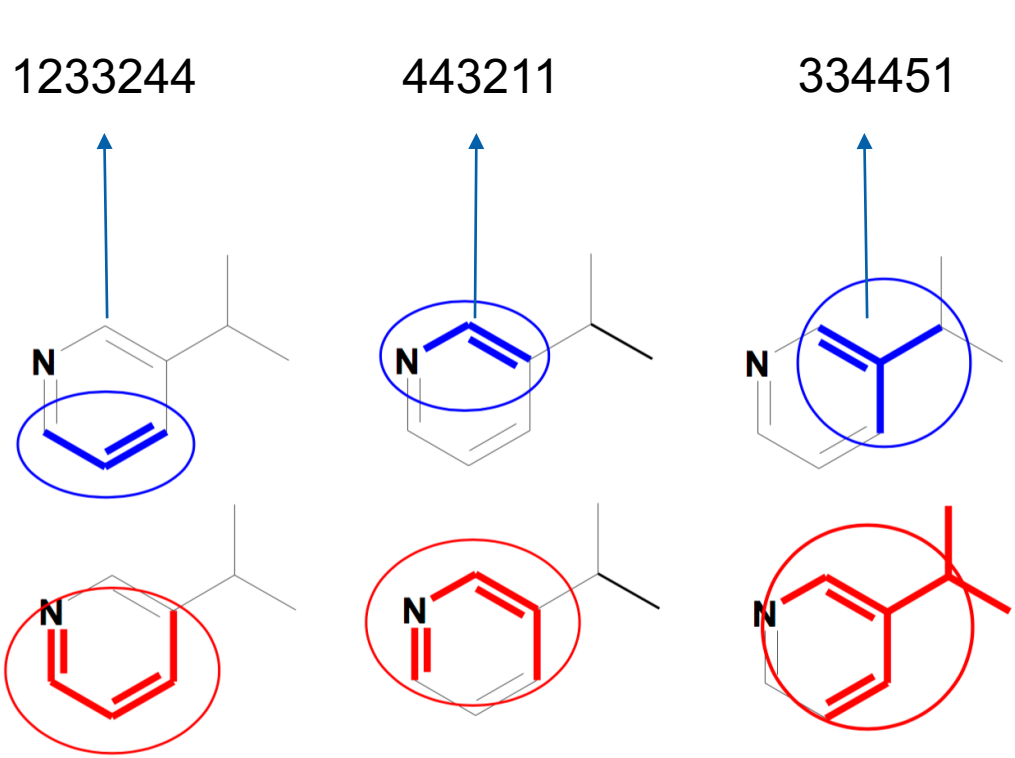
- Auto detect atom maps from Rgroups (ChemDraw/ICM)
- Auto convert dummy atoms to RGroups
- Attempt to add aromaticity to MD Files for reaction searching

New RDKit Reaction tools

Enumerate – enumeration class for enumeration and sampling

- Different sampling strategies can be used
  - *ALL* – uses current strategy
  - *RandomSample* – standard random sampling
  - *RandomSampleAllBBs* – enforces sampling of all reagents
  - *EvenSamplePairs* – useful for sampling a small number of products trying to use as many pairs of reagents as possible
- Picklable
  - pickle and restore building blocks and reaction
- Restartable
  - get nth sample, save state and continue later.

QSAR in enumeration space



Atoms
"hashed"
Into integer

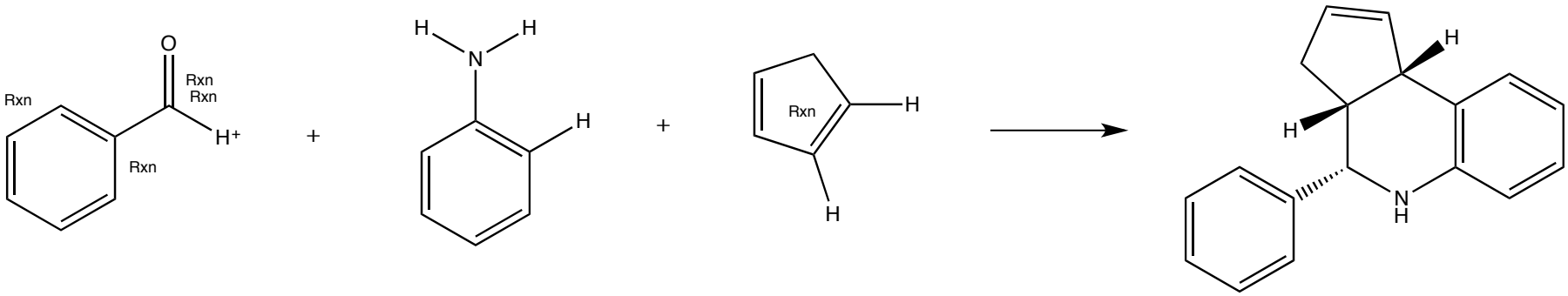
radius=1

Almost Additive

radius=2

Somewhat Additive

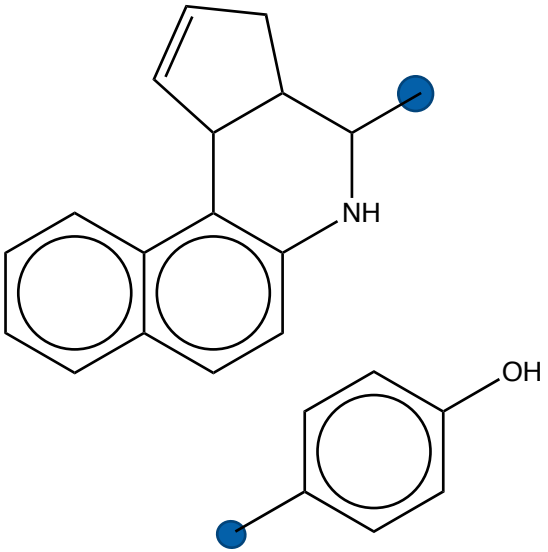
Additive fingerprints



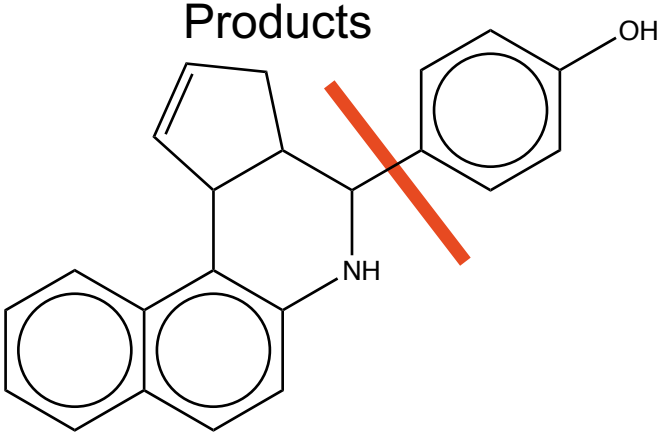
Grieco three component condensation

Additive fingerprints

Reagents



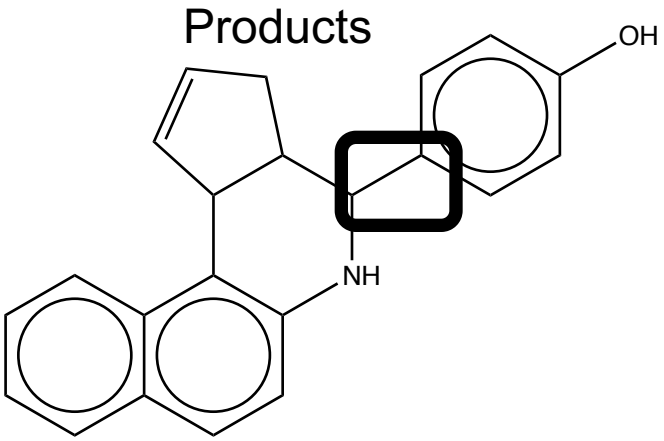
Products



Self similarity by adding bits
.7 (radius = 1)
.6 (radius = 2)

Use information you have

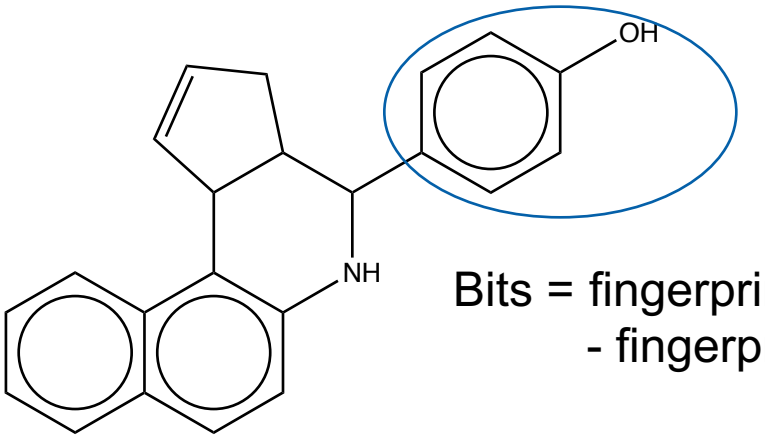
Must be C-a bond
Either: C-C c-C or C-c
Separate reactions that make different environments to add bits that *must* exist



Self similarity by adding bits
.9 (radius = 1)
.8 (radius = 2)

Search Strategy

Generate fps for each reagent in the context of the reaction

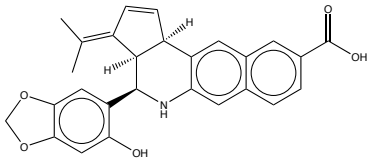


Bits = fingerprint of regeant+scaffold
- fingerprint of scaffold

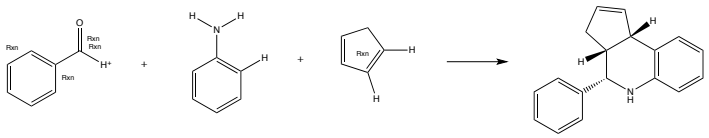
Sort reagents by similarity to target, choose top N to fully enumerate.

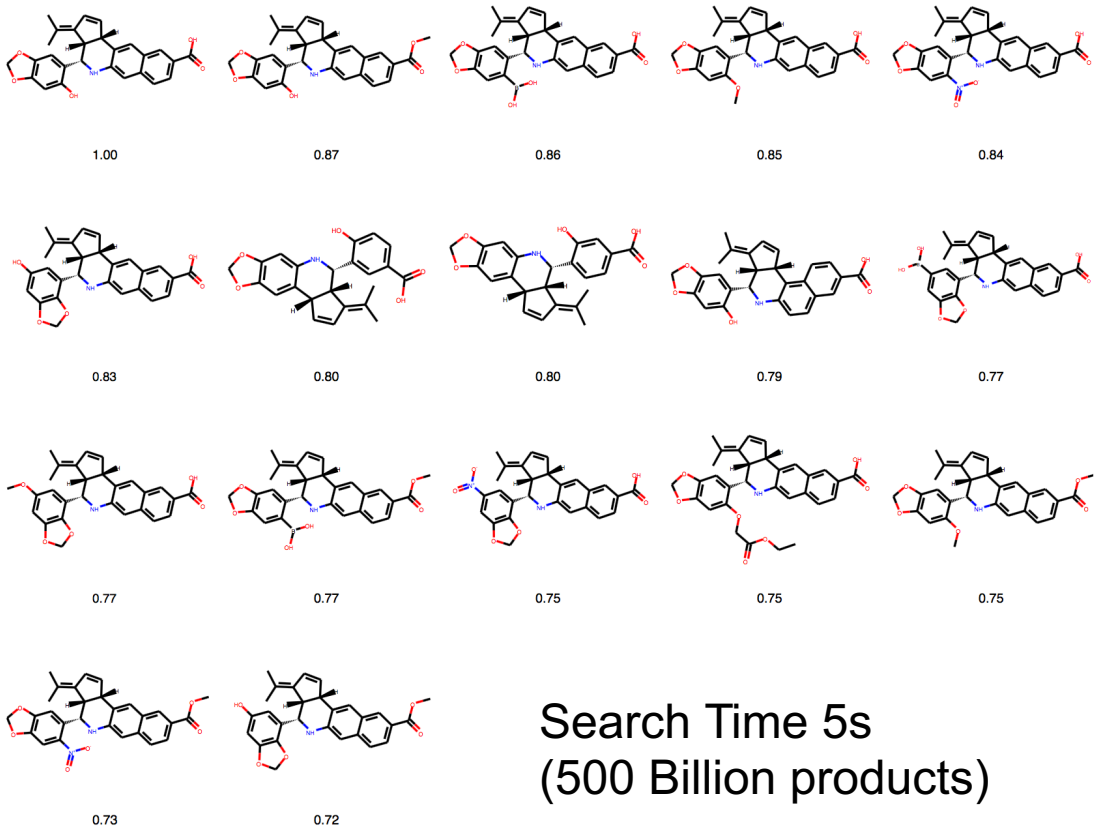
Search Strategy

query



Grieco three component condensation





Search Time 5s
(500 Billion products)

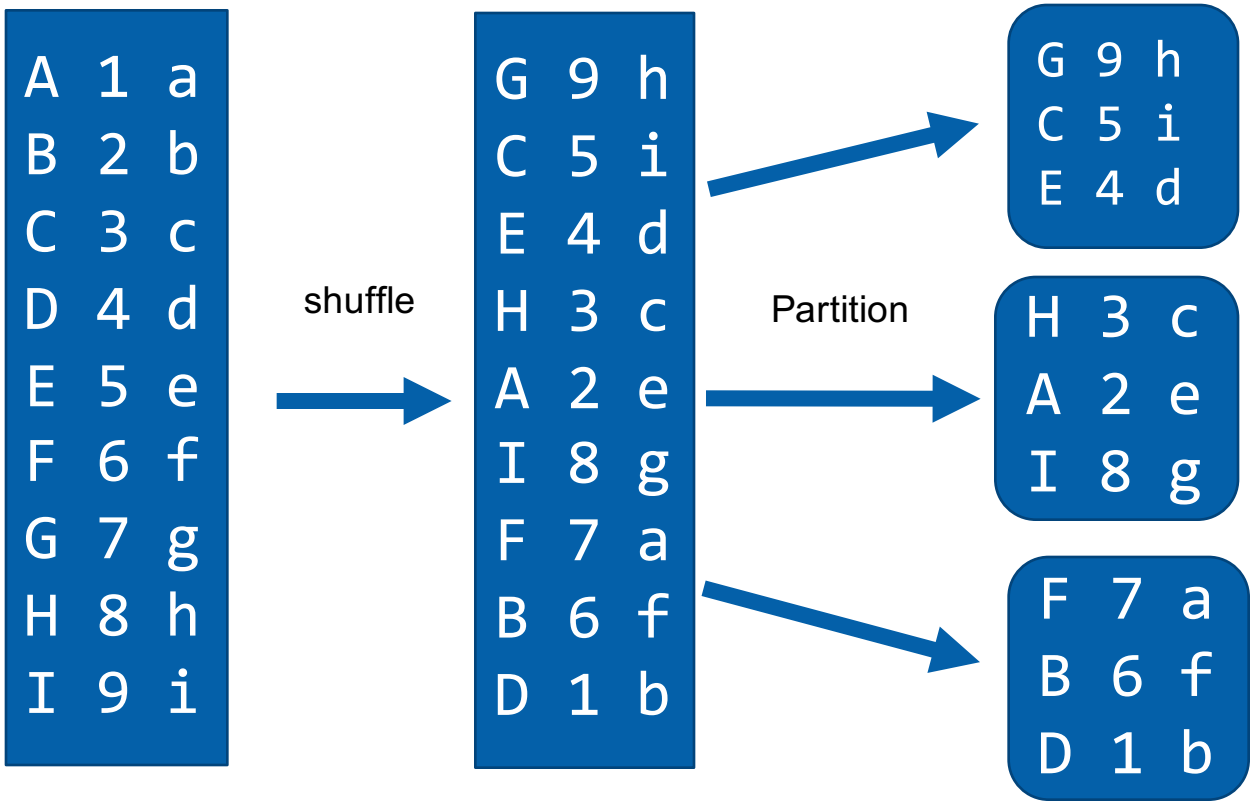
Glare: or where are the class-m(olecule) planets

Some Properties are easier than others

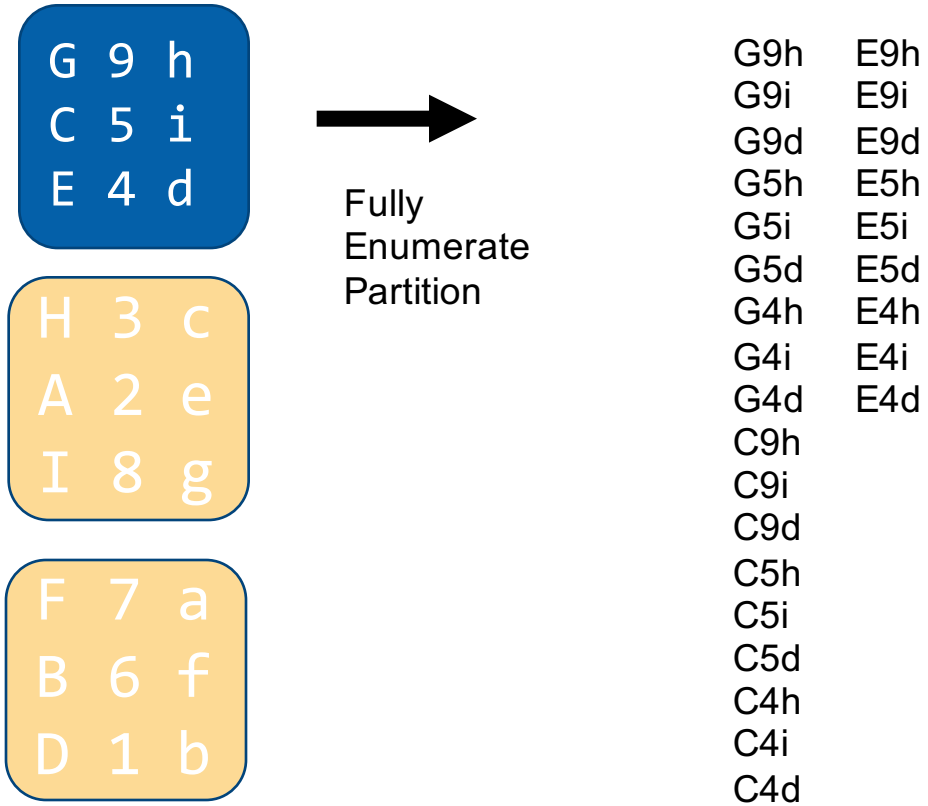
- LDC Criteria
  - MW < 650
  - CLOGP < 6.5
  - PSA < 170
  - ROTORS < 10
  - UNDEFINED STEREO CENTERS <=4

And Fingerprints are (kinda) additive as well.

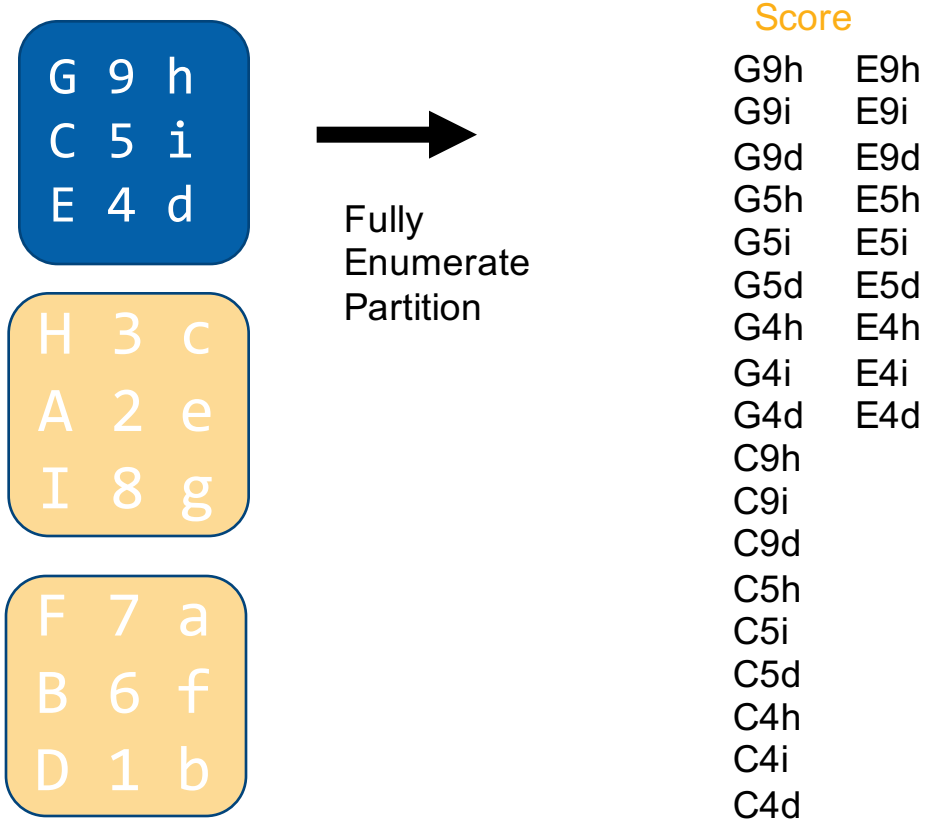
Glare


A 1 a
B 2 b
C 3 c
D 4 d
E 5 e
F 6 f
G 7 g
H 8 h
I 9 i
shuffle
G 9 h
C 5 i
E 4 d
H 3 c
A 2 e
I 8 g
F 7 a
B 6 f
D 1 b
Partition
G 9 h
C 5 i
E 4 d
H 3 c
A 2 e
I 8 g
F 7 a
B 6 f
D 1 b

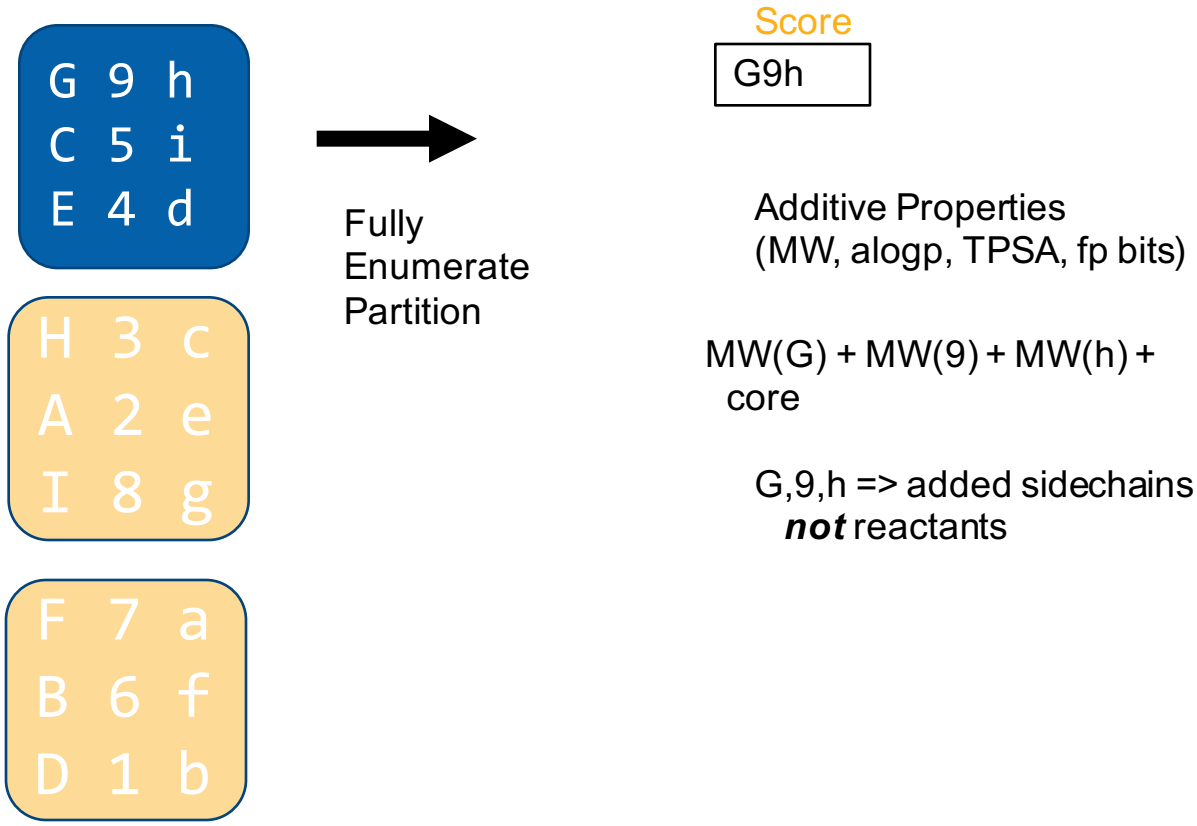
Glare


G 9 h
C 5 i
E 4 d
H 3 c
A 2 e
I 8 g
F 7 a
B 6 f
D 1 b
Fully
Enumerate
Partition
G9h
G9i
G9d
G5h
G5i
G5d
G4h
G4i
G4d
C9h
C9i
C9d
C5h
C5i
C5d
C4h
C4i
C4d
E9h
E9i
E9d
E5h
E5i
E5d
E4h
E4i
E4d

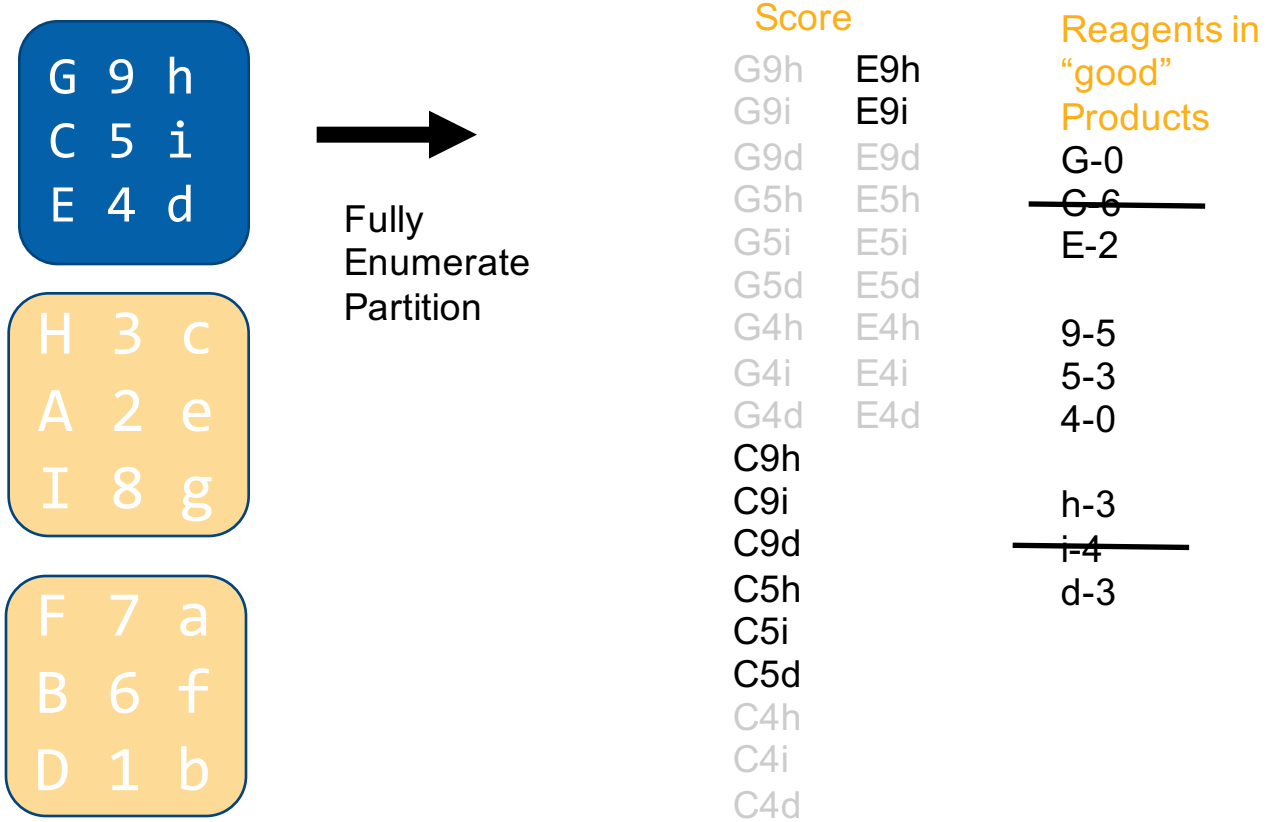
Glare


G 9 h
C 5 i
E 4 d
H 3 c
A 2 e
I 8 g
F 7 a
B 6 f
D 1 b
Fully
Enumerate
Partition
Score
G9h
G9i
G9d
G5h
G5i
G5d
G4h
G4i
G4d
C9h
C9i
C9d
C5h
C5i
C5d
C4h
C4i
C4d
E9h
E9i
E9d
E5h
E5i
E5d
E4h
E4i
E4d

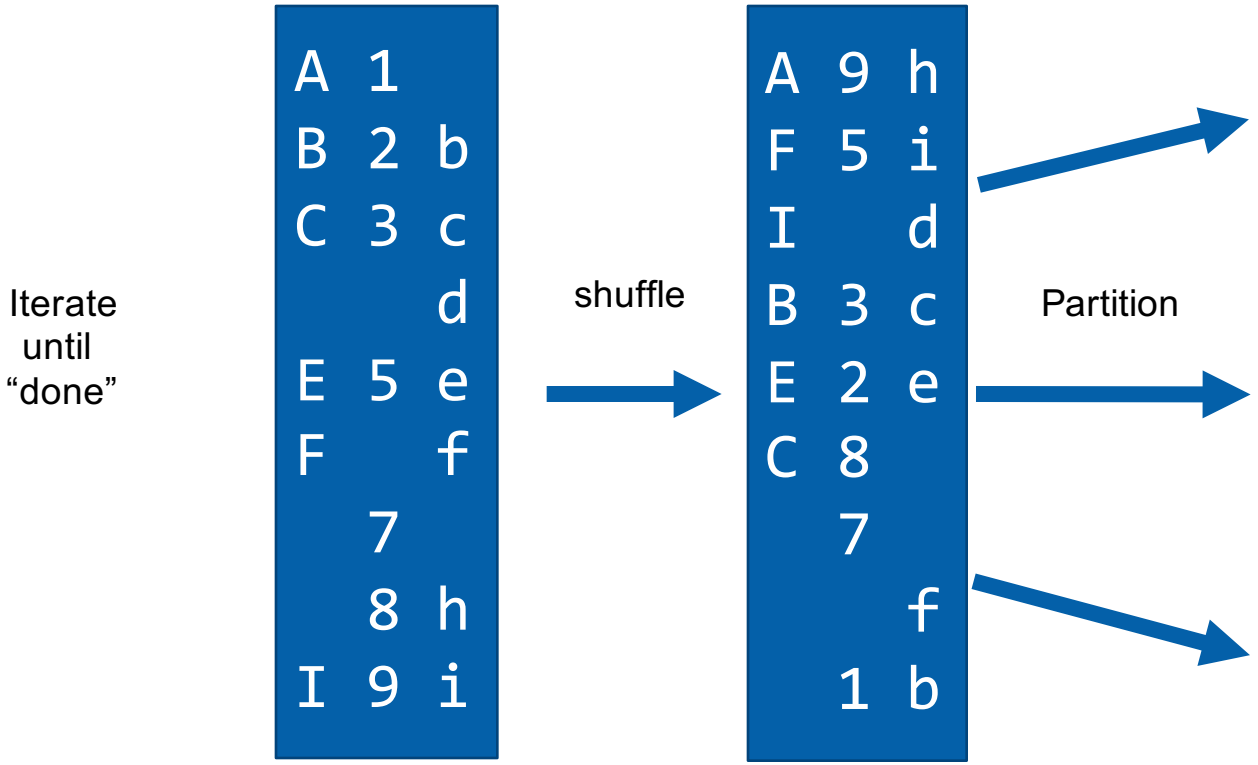
Glare


G 9 h
C 5 i
E 4 d
H 3 c
A 2 e
I 8 g
F 7 a
B 6 f
D 1 b
Fully
Enumerate
Partition
Score
G9h
Additive Properties
(MW, alogp, TPSA, fp bits)
MW(G) + MW(9) + MW(h) +
core
G,9,h => added sidechains
not reactants

Glare


G 9 h
C 5 i
E 4 d
H 3 c
A 2 e
I 8 g
F 7 a
B 6 f
D 1 b
Fully
Enumerate
Partition
Score
G9h
G9i
G9d
G5h
G5i
G5d
G4h
G4i
G4d
C9h
C9i
C9d
C5h
C5i
C5d
C4h
C4i
C4d
E9h
E9i
E9d
E5h
E5i
E5d
E4h
E4i
E4d
Reagents in "good" Products
G-0
C-6
E-2
9-5
5-3
4-0
h-3
i-4
d-3

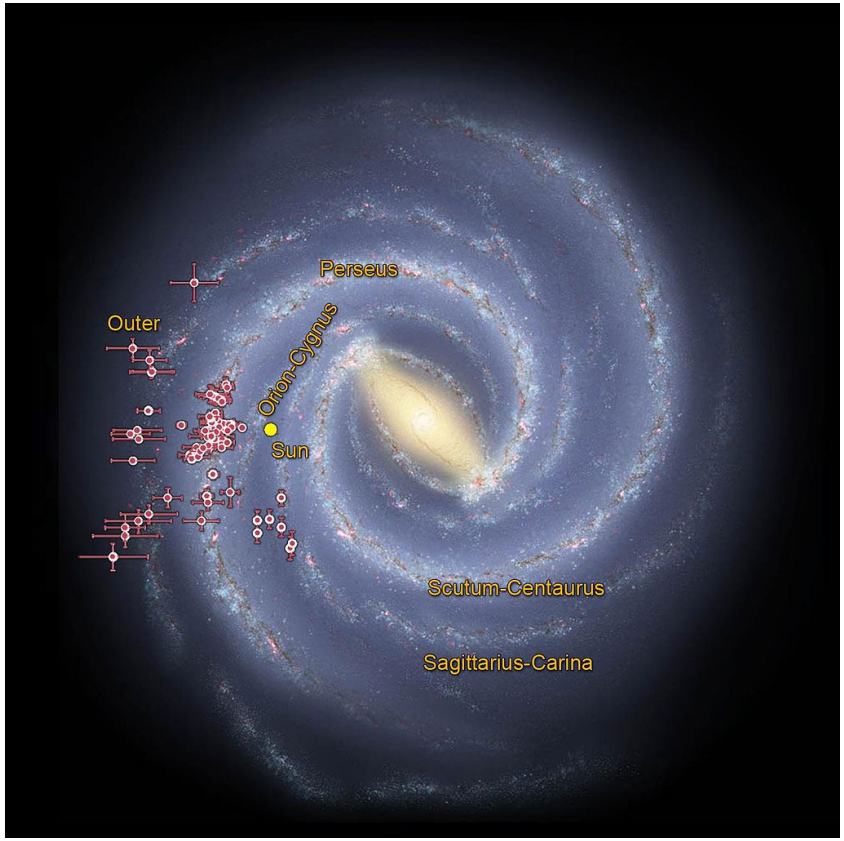
Glare


Iterate
until
“done”
A 1
B 2 b
C 3 c
d
E 5 e
F f
7
8 h
I 9 i
shuffle
A 9 h
F 5 i
I d
B 3 c
E 2 e
C 8
7
f
1 b
Partition

Glare

Takes about 1 minute to search 200 billion products

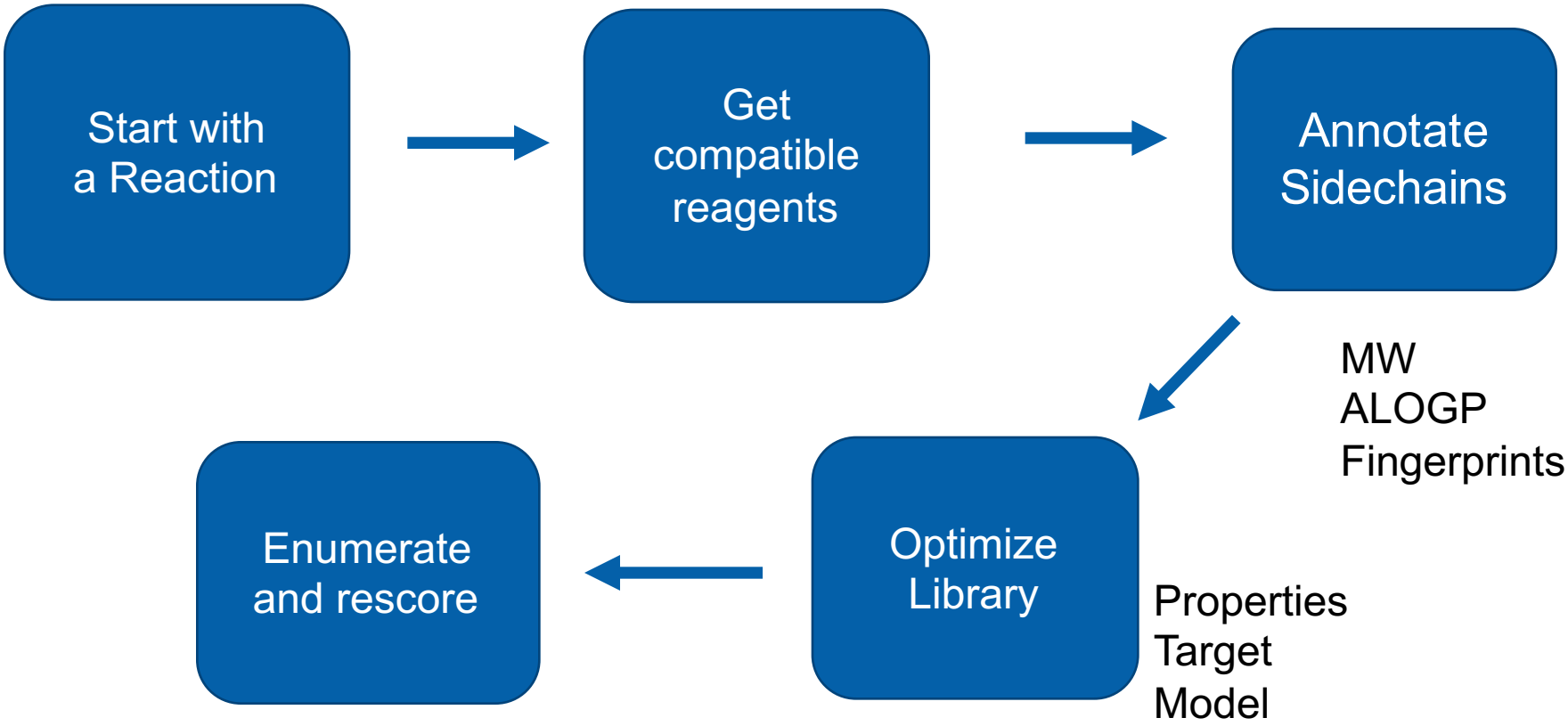
- 400 lines of python
- Including comments



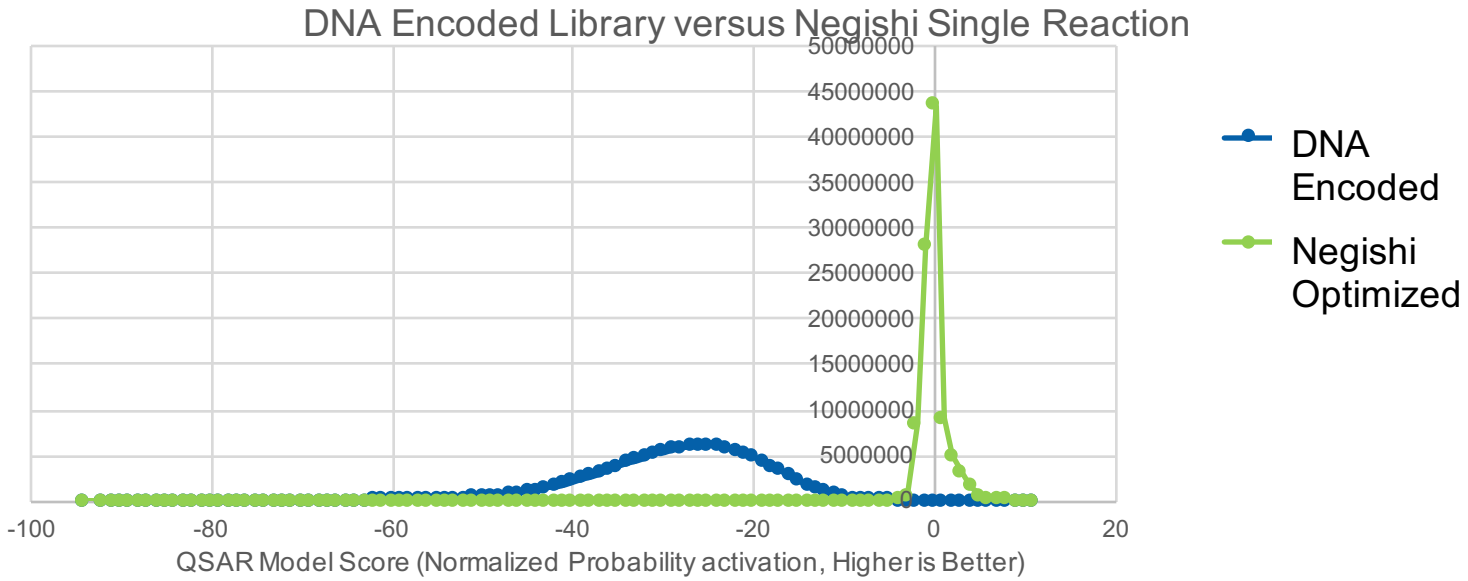
https://commons.wikimedia.org/wiki/File:PIA19341-MilkyWayGalaxy-SpiralArmsData-WISE-20150603.jpg

Truchon. Bayly. Jchem Inf Model 2006 Jul-Aug;46(4):1536-48. GLARE: a new approach for filtering large reagent lists in combinatorial library design using product properties.

QSAR in reaction space



Finding needles in haystacks



DNA Encoded Library ~ 150M compounds

Single Reaction synthesis is “top” 150 million out of ~200 billion

– More variety in Negishi Reaction reagents.

Finding needles in haystacks

Not a be all end all

- Need the reactions
- Need the reagents
- Puts a **lot** of pressure on your models.

But…

- Fast (at least get the wrong answers quicker)
- Can now use QSAR models to help choose reagent diversity
- When dealing with trillions of products…

Acknowledgements

Aileen Novero (now at Vertex)

Gregory Landrum (RDKit/Knime)

Nik Stiefl (Novartis)

Clayton Springer (Novartis)

Andrew Dalke (ChemFP)

Michael Tarselli (Novartis)

Margaret Pancost-Heidebrecht (Novartis)

Thank you

NOVARTIS