2nd March 2022

COMPUTER AIDED SYNTHESIS PREDICTION TO ENABLE AUGMENTED CHEMICAL DISCOVERY AND CHEMICAL SPACE EXPLORATION

DR AMOL THAKKAR

amol.thakkar@unibe.ch

UNIVERSITÄT BERN

MARIE CURIE ACTIONS

BigChem

AstraZeneca

INTRODUCTIONS

2012-2017
University of St Andrews, MChem, First Class Honors
Thesis: Modelling Uranyl Chemistry in Liquid Ammonia From Density Functional Theory

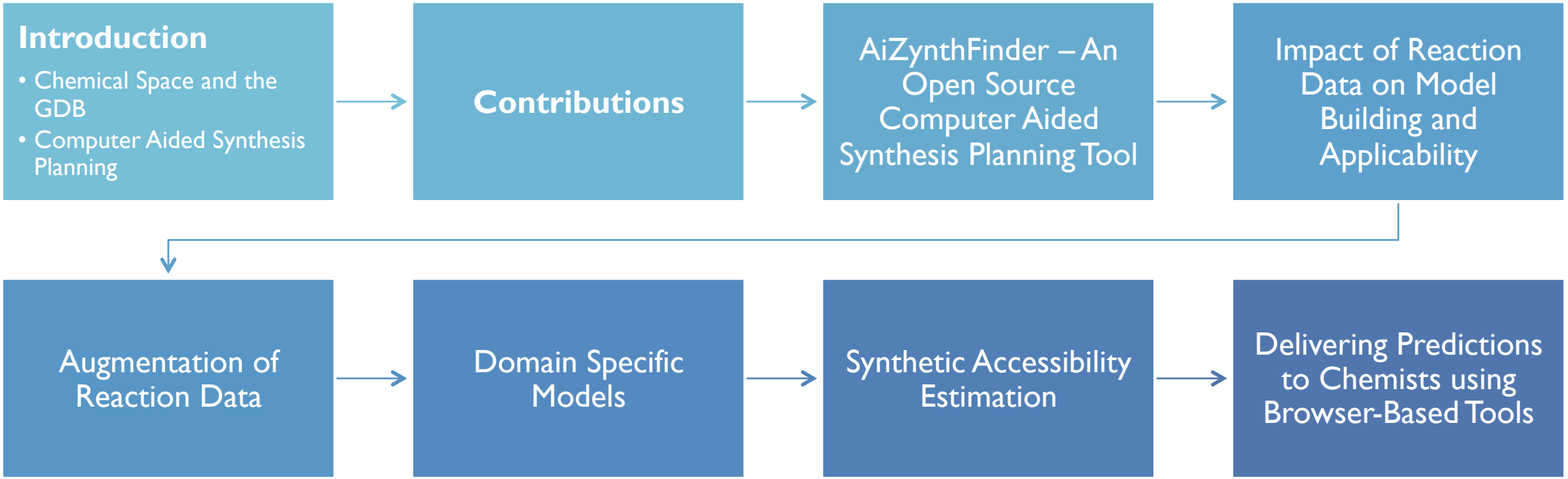
2015-2016
Pfizer, Chemical Research and Development
Process route souting, design, and optimisation, reaction screening, multi-gram synthesis, technology transfer

2018 – 2022
University of Bern, Reymond Group
Exploration of Chemical Space by Data and AI driven Computer Aided Synthesis Planning

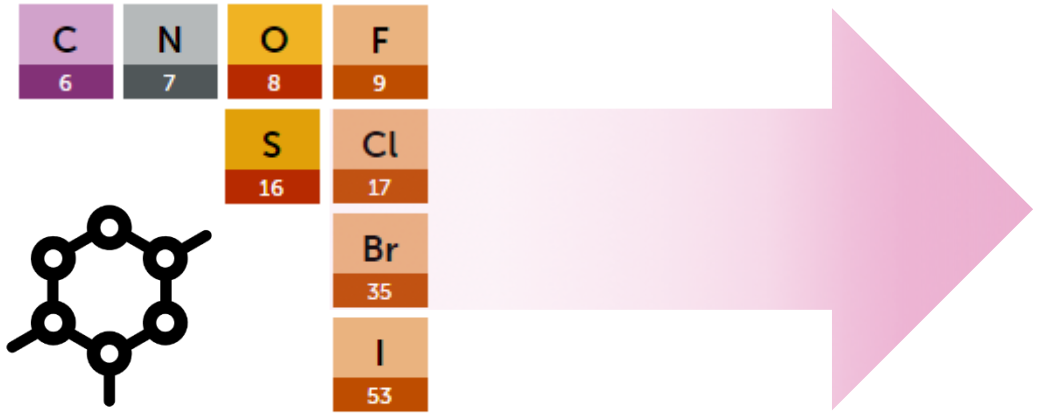
January 2019 - July 2020
AstraZeneca, Molecular Artificial Intelligence

- Co-developed and led efforts for a computer aided synthesis planning platform that was deployed in industry, combining automated data extraction from inhomogeneous sources, artificial neural networks, and Monte-Carlo tree search.
- Used the developed tools to identify the synthetic accessibility of chemical space, and impacted ongoing drug discovery and development projects.

OVERVIEW


Introduction
• Chemical Space and the GDB
• Computer Aided Synthesis Planning
Contributions
AiZynthFinder – An Open Source Computer Aided Synthesis Planning Tool
Impact of Reaction Data on Model Building and Applicability
Augmentation of Reaction Data
Domain Specific Models
Synthetic Accessibility Estimation
Delivering Predictions to Chemists using Browser-Based Tools

CHEMICAL SPACE AND **DRUG DISCOVERY**

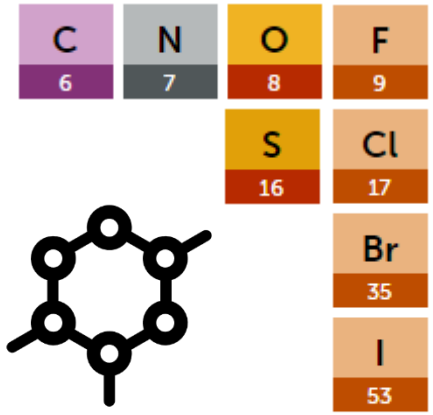


Organic Chemistry

10^{60} **Molecules**
Orally Bioavailable
Obeying Lipinskis Rule of 5

Bohacek et al. Med. Res. Rev. ***1996****, 16, 3– 50*
Lipinski et al. Adv. Drug Delivery Rev. ***1997****, 23, 3–25*

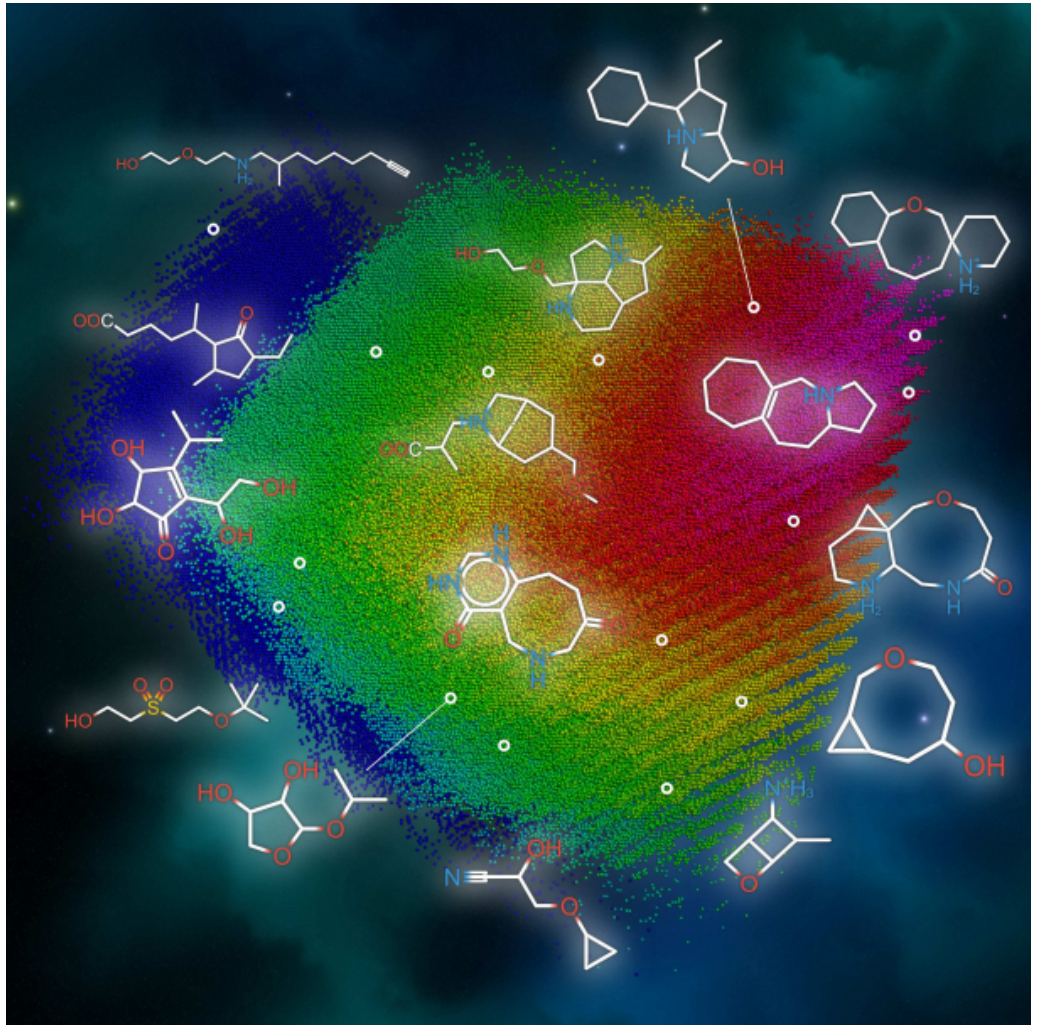
CHEMICAL SPACE AND **GDB**



ca. 166 billion

GDB17

Ruddigkeit et al., *J. Chem. Inf. Model.* **2012**, 52, 11, 2864–2875



KEY RESEARCH QUESTION

How do we **predict** the **synthesis** of **compounds** that are **generated** or **suggested** by a **chemist or computer**?

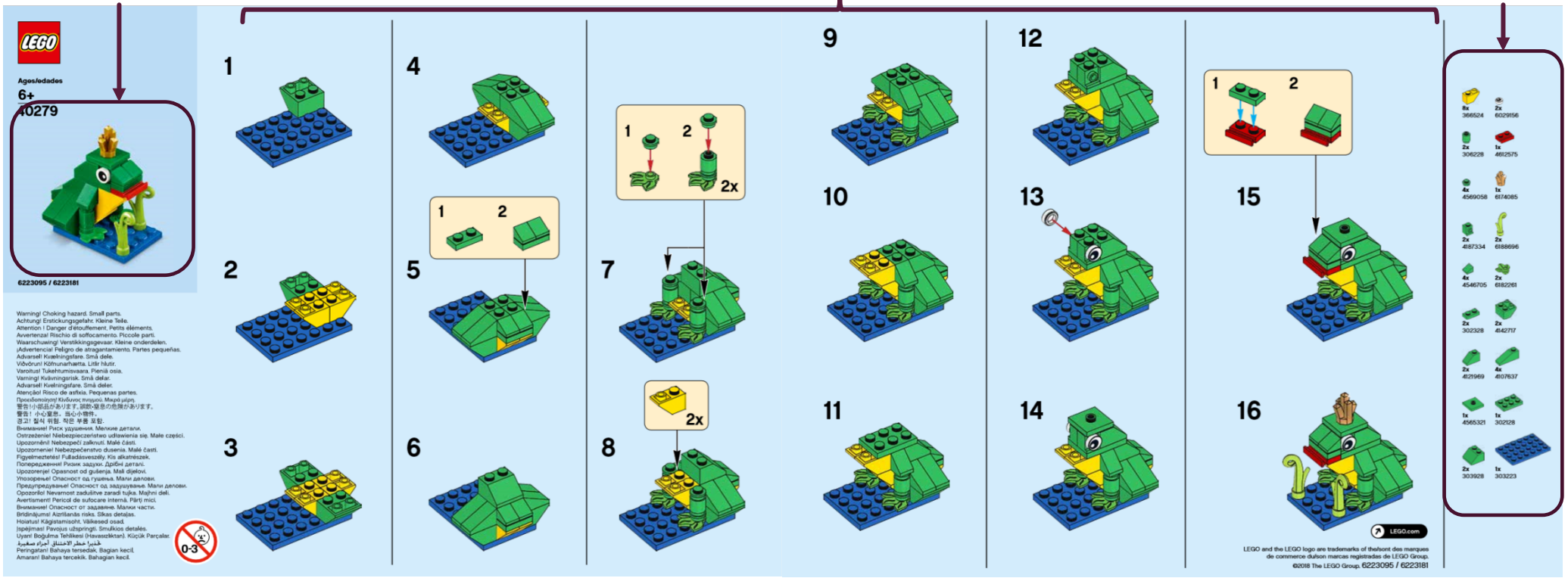
LEGO – ANALOGY TO SYNTHESIS PLANNING

Building Lego as an analogy to synthetic route planning

Target Molecule

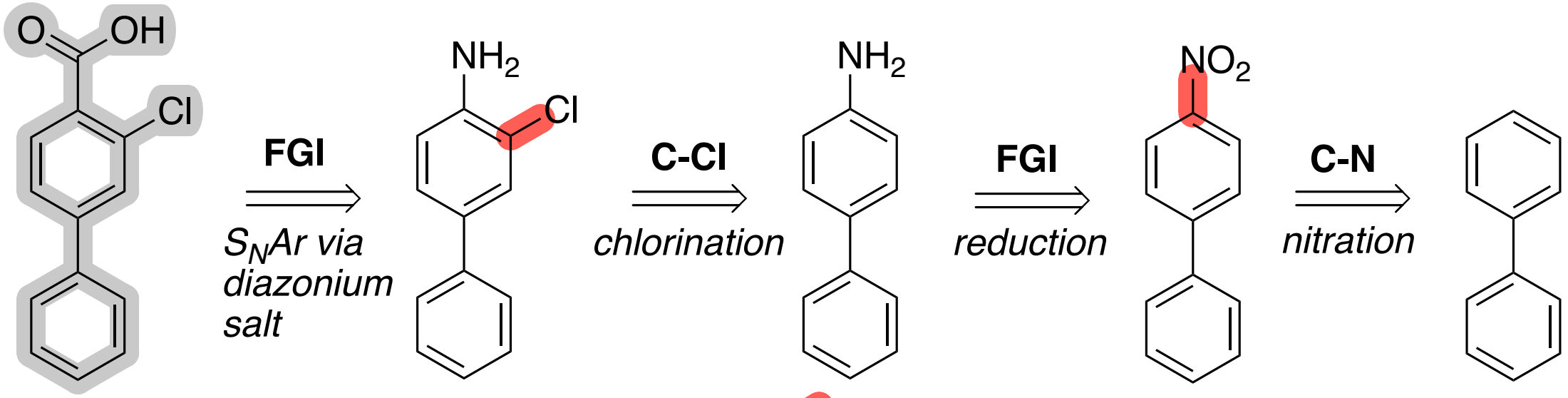
Sequence of steps to build target

Known Building Blocks



SYNTHESIS PLANNING

Retrosynthetic Analysis



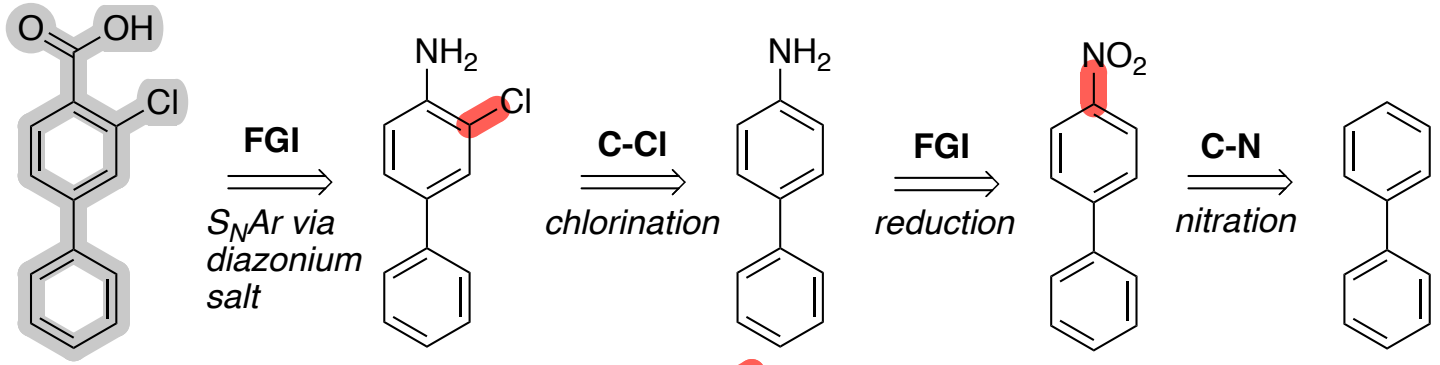
FGI = Functional Group Interconversion

Disconnection = Bond Breaking/Forming Step

Byron et. al. J. Chem. Soc. 1963, 2253
Byron et. al. J. Chem. Soc. 1966, 840
Warren and Wyatt. Organic Synthesis: The Disconnection Approach, Wiley, 2011

SYNTHESIS PLANNING

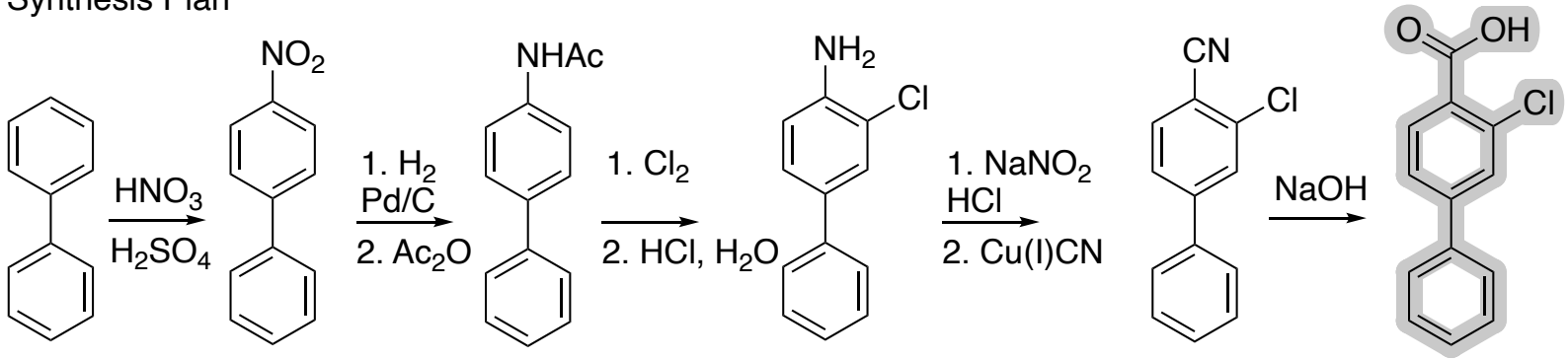
Retrosynthetic Analysis



FGI = Functional Group Interconversion

Disconnection = Bond Breaking/Forming Step

Synthesis Plan



Byron et. al. J. Chem. Soc. 1963, 2253
Byron et. al. J. Chem. Soc. 1966, 840
Warren and Wyatt. Organic Synthesis: The Disconnection Approach, Wiley, 2011

COMPUTER AIDED SYNTHESIS PLANNING (CASP)

1

LHASA—Logic and Heuristics Applied to Synthetic Analysis

DAVID A. PENSAK

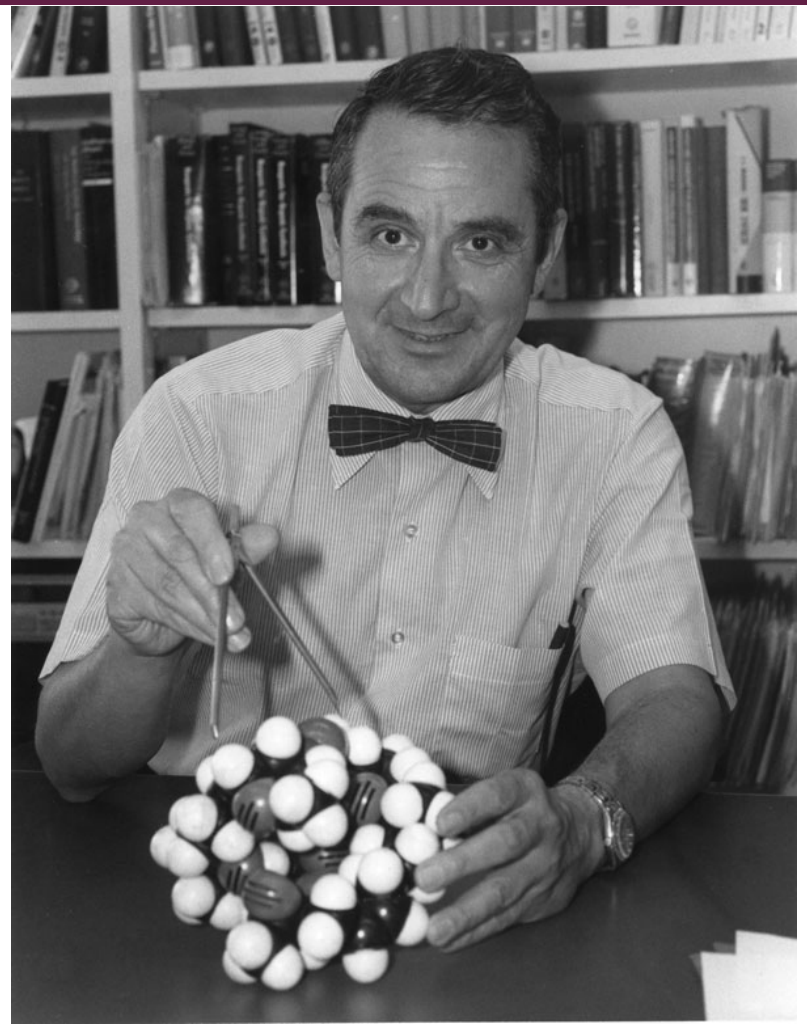
Central Research and Develop. Dept., E. I. du Pont de Nemours and Co., Wilmington, Del. 19898

E. J. COREY

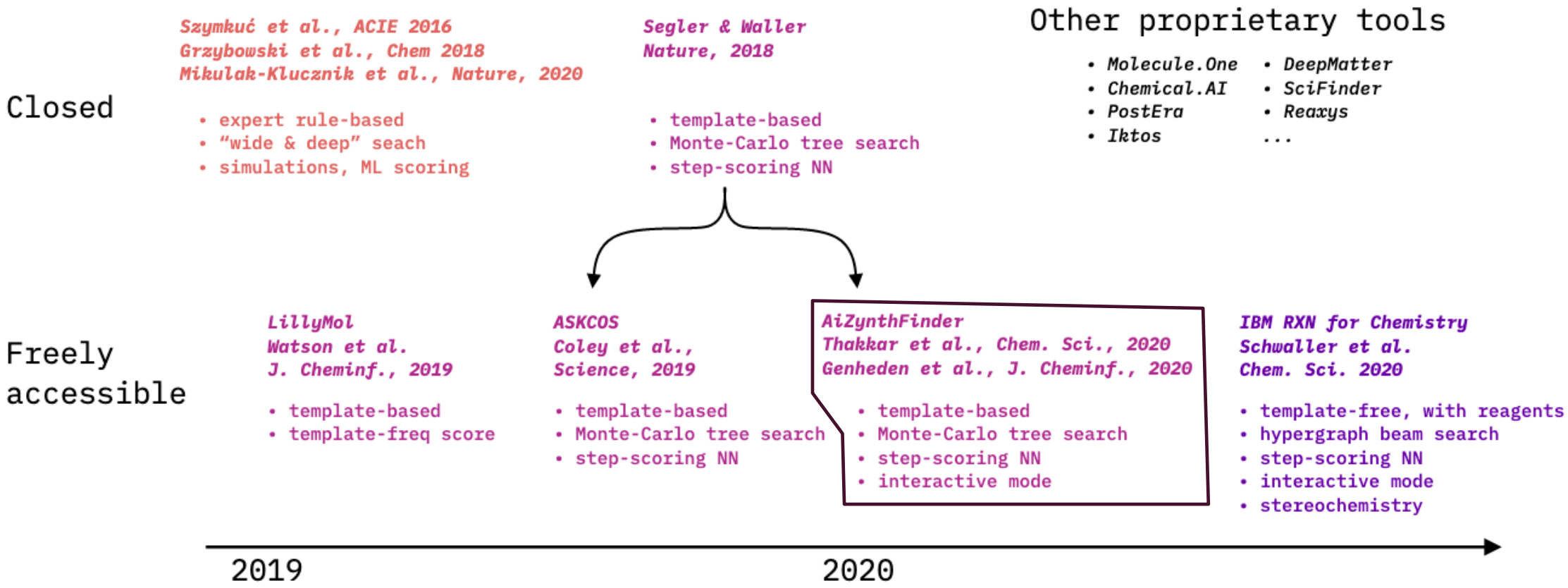
Dept. of Chemistry, Harvard University, Cambridge, Mass. 02138

Despite the wealth of knowledge about various chemical reactions, there exists no formal framework of interrelationships to guide the chemist in the synthesis of even moderately complex molecules. The LHASA (Logic and Heuristics Applied to Synthetic Analysis) project is an attempt to codify and organize the techniques used in organic synthesis.

One important aspect of the project has been the writing of a general purpose computer program which will aid the laboratory chemist and will employ both the basic and more complex techniques for synthetic design as elucidated by this study. The program (hereafter also called LHASA) is intended to propose a variety of synthetic routes to whatever molecule it is given. The responsibility for final evaluation of the merit of the routes lies with the chemist. The

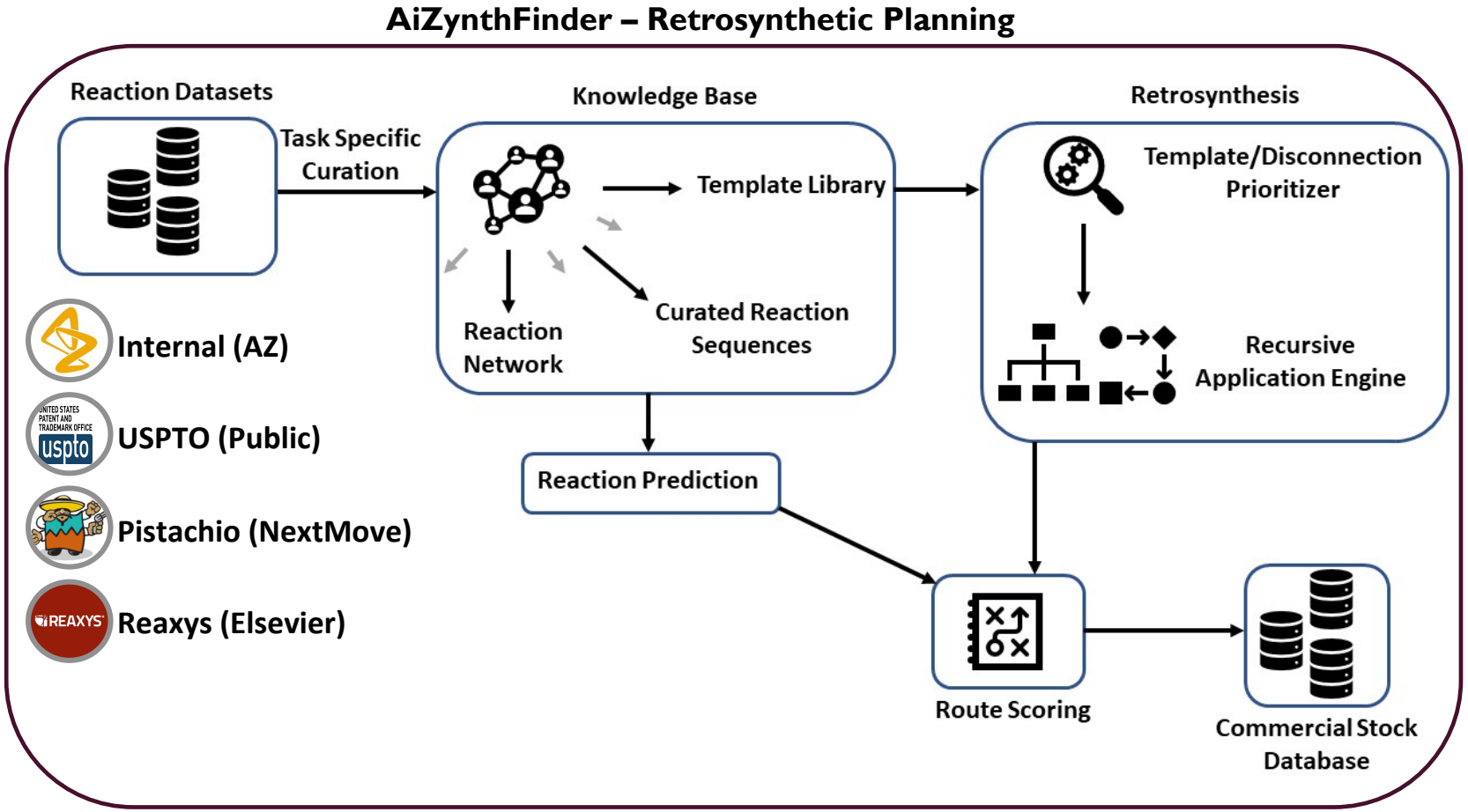


MODERN CASP



Philippe Schwaller PhD Thesis, 2021, University of Bern

DEVELOPING AIZYNTHFINDER



Thakkar A et al., *Chemical Science*, **2020**, *11* (1), 154-168.
Genheden S, Thakkar A et al., *J. Cheminform.*, **2020**, *12:70*

Highlights:
Most popular 2019-2020 physical and theoretical chemistry articles and Accelerating Chemistry Symposium Collection

REACTION DATA - INCONSISTENCIES

Dataset

Proprietary

Internal (AZ)

Literature

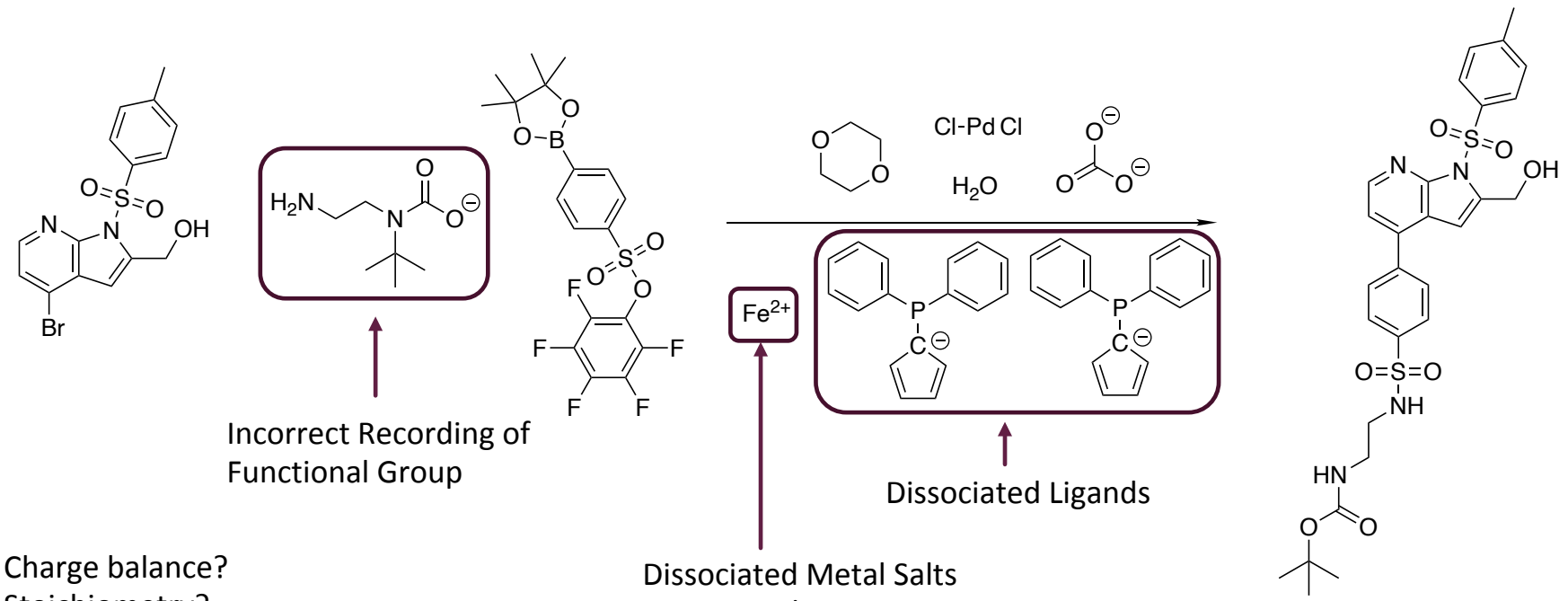
Reaxys (Elsevier)

Patents

USPTO (Public)

Pistachio (NextMove)

Example from USPTO

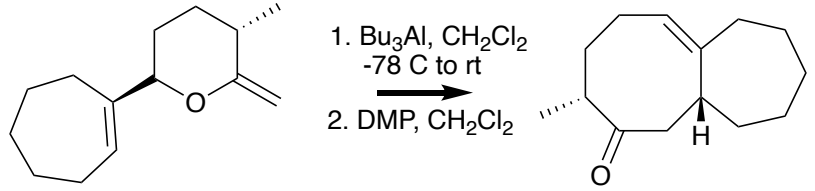


Charge balance?
Stoichiometry?
Conditions missing from meta-data
Multi-step reactions

Thakkar et al., *Chemical Science*, **2020,** *11* (1), 154-168.

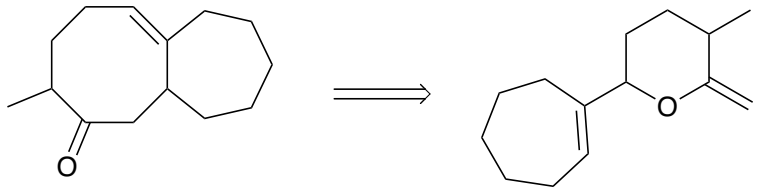
TEMPLATE EXTRACTION

Reaction Example – Claisen Rearrangement



Reactant **Product**

Retro Reaction Example – Claisen Rearrangement

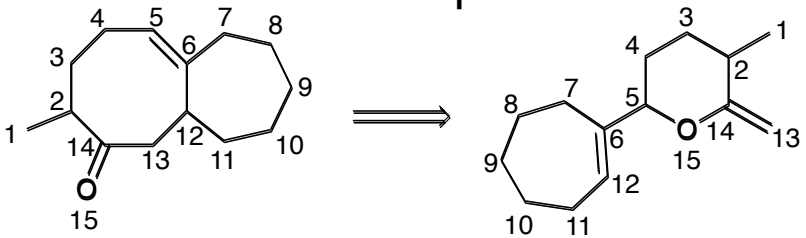


Product **Reactant**

Thakkar et al., *Chemical Science,* **2020,** *11* (1), 154-168.
Coley et al. *Journal of Chemical Information and Modelling* **2019**, *59* (6), 2529–2537.
Sun et al. *Chem. – Asian J.* **2012**, 7, 2321

TEMPLATE EXTRACTION

Retro Reaction Example – Claisen Rearrangement



Reaction SMILES

CC1CC/C=C2C(CC1=O)CCCCC\2>>CC3CCC(OC3=C)C4=CCCCCC4

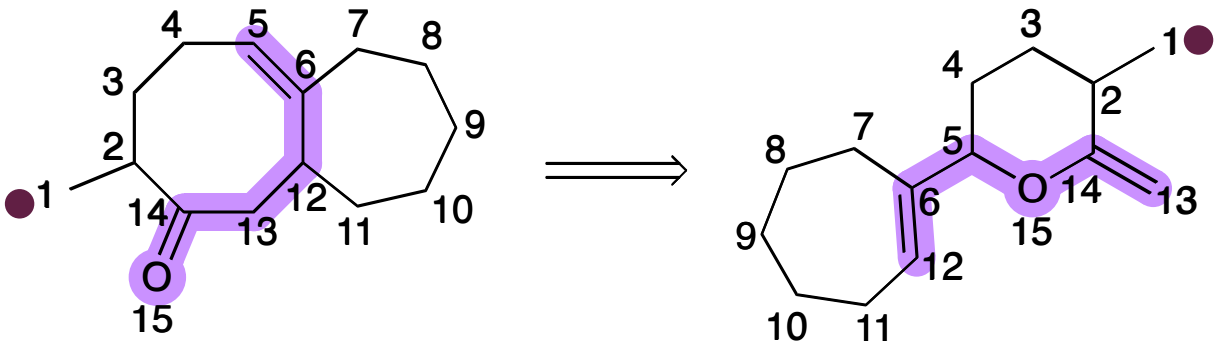
Atom-mapped Reaction SMILES

[CH3:1][CH:2]1[CH2:3][CH2:4][CH:5]([C:6]2=[CH:12][CH2:11][CH2:10][CH2:9][CH2:8][CH2:7]2)[O:15][C:14]1=[CH2:13]>>[CH3:1][CH:2]1[CH2:3][CH2:4]/[CH:5]=[C:6]2/[CH2:7][CH2:8][CH2:9][CH2:10][CH2:11]CH:12]2[CH2:13][C:14]1=[O:15]

Thakkar et al., *Chemical Science,* **2020,** *11* (1), 154-168.
Coley et al. *Journal of Chemical Information and Modelling* **2019**, *59* (6), 2529–2537.
Sun et al. *Chem. – Asian J.* **2012**, 7, 2321

TEMPLATE EXTRACTION

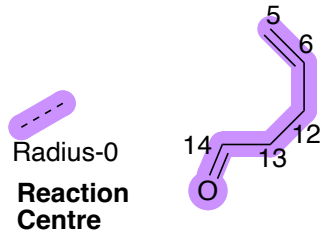
Retro Reaction Example – Claisen Rearrangement



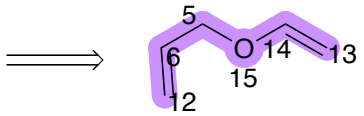
Procedure

- Iterate around atom indices
- Check for a change in atomic environment
- All changes = reaction centre
- Extract reaction centre as template (Reaction SMARTS)
- Expand Core

Product



Reactant



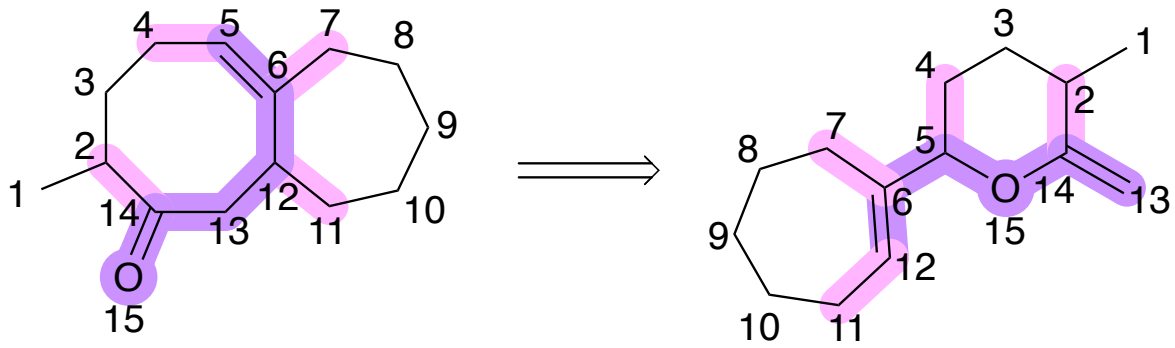
Reaction SMARTS – Template Radius 0

([CH;D2;+0:4]=[C;H0;D3;+0:5]\\[CH;D3;+0:6]-[CH2;D2;+0:1]-[C;H0;D3;+0:2]=[O;H0;D1;+0:3])>>([CH2;D1;+0:1]=[C;H0;D3;+0:2]-[O;H0;D2;+0:3]-[CH;D3;+0:4]-[C;H0;D3;+0:5]=[CH;D2;+0:6])

Thakkar et al., *Chemical Science*, **2020,** *11* (1), 154-168.
Coley et al. *Journal of Chemical Information and Modelling* **2019**, *59* (6), 2529–2537.
Sun et al. *Chem. – Asian J.* **2012**, 7, 2321

TEMPLATE EXTRACTION

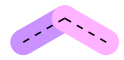
Retro Reaction Example – Claisen Rearrangement

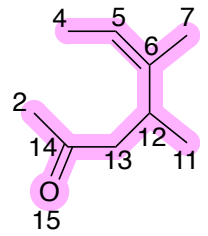


Core Expansion

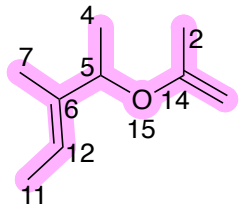
- Consider Neighbouring Atoms
- Increased Template Specificity
- Increases Number of Templates Extracted

Product





Reactant



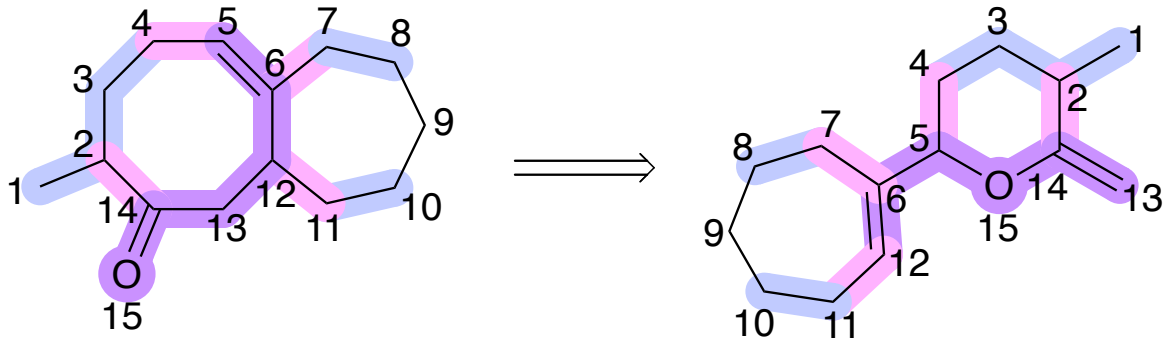
Reaction SMARTS – Template Radius 1

([C:1]-[CH;D3;+0:2](-[CH2;D2;+0:10]-[C;H0;D3;+0:8](-[C:9])=[O;H0;D1;+0:7])/[C;H0;D3;+0:3](-[C:4])=[CH;D2;+0:5]\\[C:6])>>([C:1]-[CH;D2;+0:2]=[C;H0;D3;+0:3](-[C:4])-[CH;D3;+0:5](-[C:6])-[O;H0;D2;+0:7]-[C;H0;D3;+0:8](-[C:9])=[CH2;D1;+0:10])

Thakkar et al., *Chemical Science*, **2020,** *11* (1), 154-168.
Coley et al. *Journal of Chemical Information and Modelling* **2019**, *59* (6), 2529–2537.
Sun et al. *Chem. – Asian J.* **2012**, 7, 2321

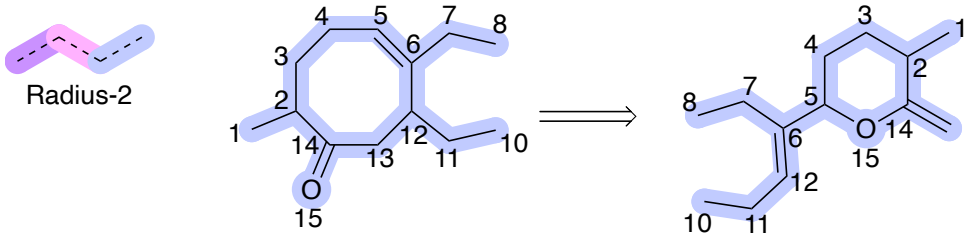
TEMPLATE EXTRACTION

Retro Reaction Example – Claisen Rearrangement



Core Expansion

- Consider Neighbouring Atoms
- Increased Template Specificity
- Increases Number of Templates Extracted

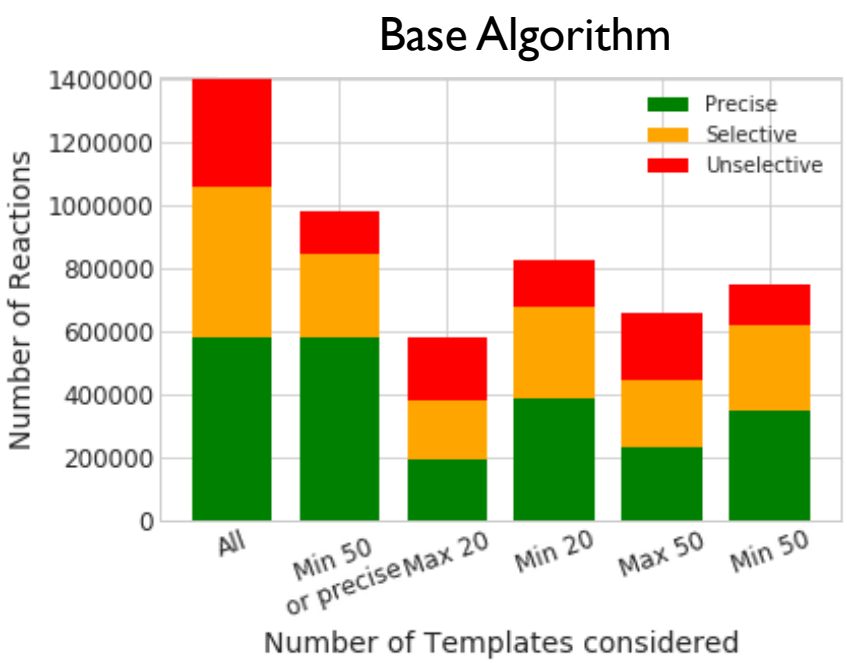


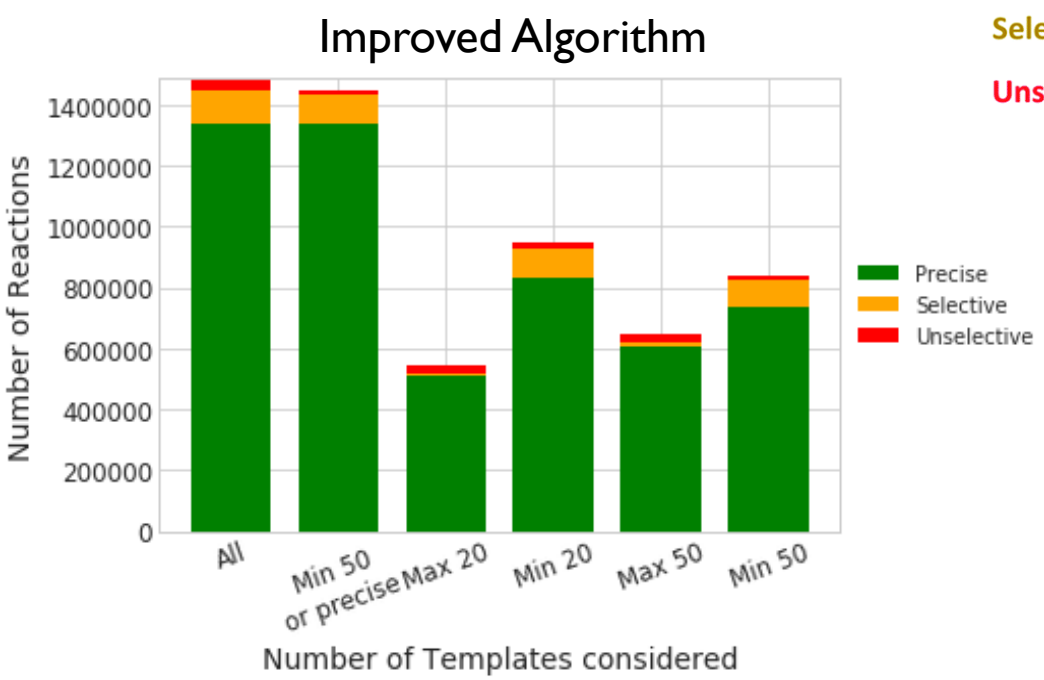
Reaction SMARTS – Template Radius 2

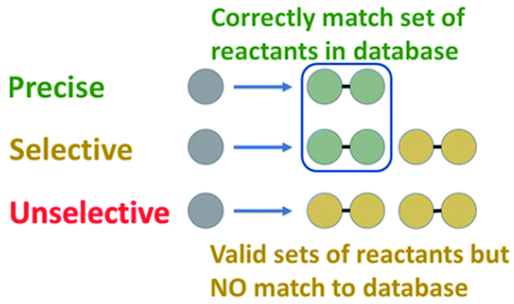
([C:1]-[C:2]-[CH;D3;+0:3]1-[CH2;D2;+0:13]-[C;H0;D3;+0:12](=[O;H0;D1;+0:14])-[C:10](-[C;D1;H3:11])-[C:9]-[C:8]/[CH;D2;+0:7]=[C;H0;D3;+0:4]\\1-[C:5]-[C:6])>>([C:1]-[C:2]-[CH;D2;+0:3]=[C;H0;D3;+0:4](-[C:5]-[C:6])-[CH;D3;+0:7]1-[C:8]-[C:9]-[C:10](-[C;D1;H3:11])-[C;H0;D3;+0:12](=[CH2;D1;+0:13])-[O;H0;D2;+0:14]-1)

Thakkar et al., *Chemical Science*, **2020,** *11* (1), 154-168.
Coley et al. *Journal of Chemical Information and Modelling* **2019**, *59* (6), 2529–2537.
Sun et al. *Chem. – Asian J.* **2012**, 7, 2321

TEMPLATE QUALITY ASSESSMENT

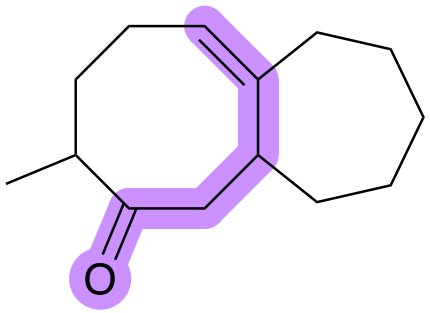






Thakkar et al., *Chemical Science*, **2020,** *11* (1), 154-168.
Coley et al. *Journal of Chemical Information and Modelling* **2019**, *59* (6), 2529–2537.

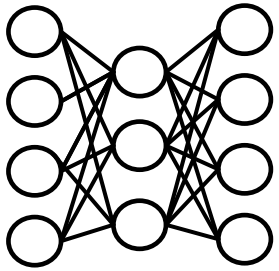
NEURAL NETWORK TRAINING – TEMPLATE PRIORITISATION



Product

Molecule Encoded as a Bit Vector

ECFP4

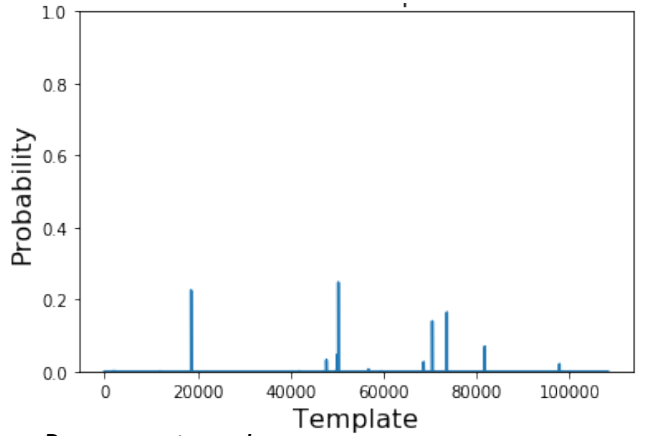


Policy Network

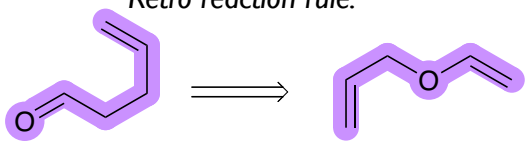
One-Hot Encoding of Templates

Template 1
Template 2
Template 3
Template 4
Template 5
Template N

Predict Retrosynthesis Template



Retro reaction rule:

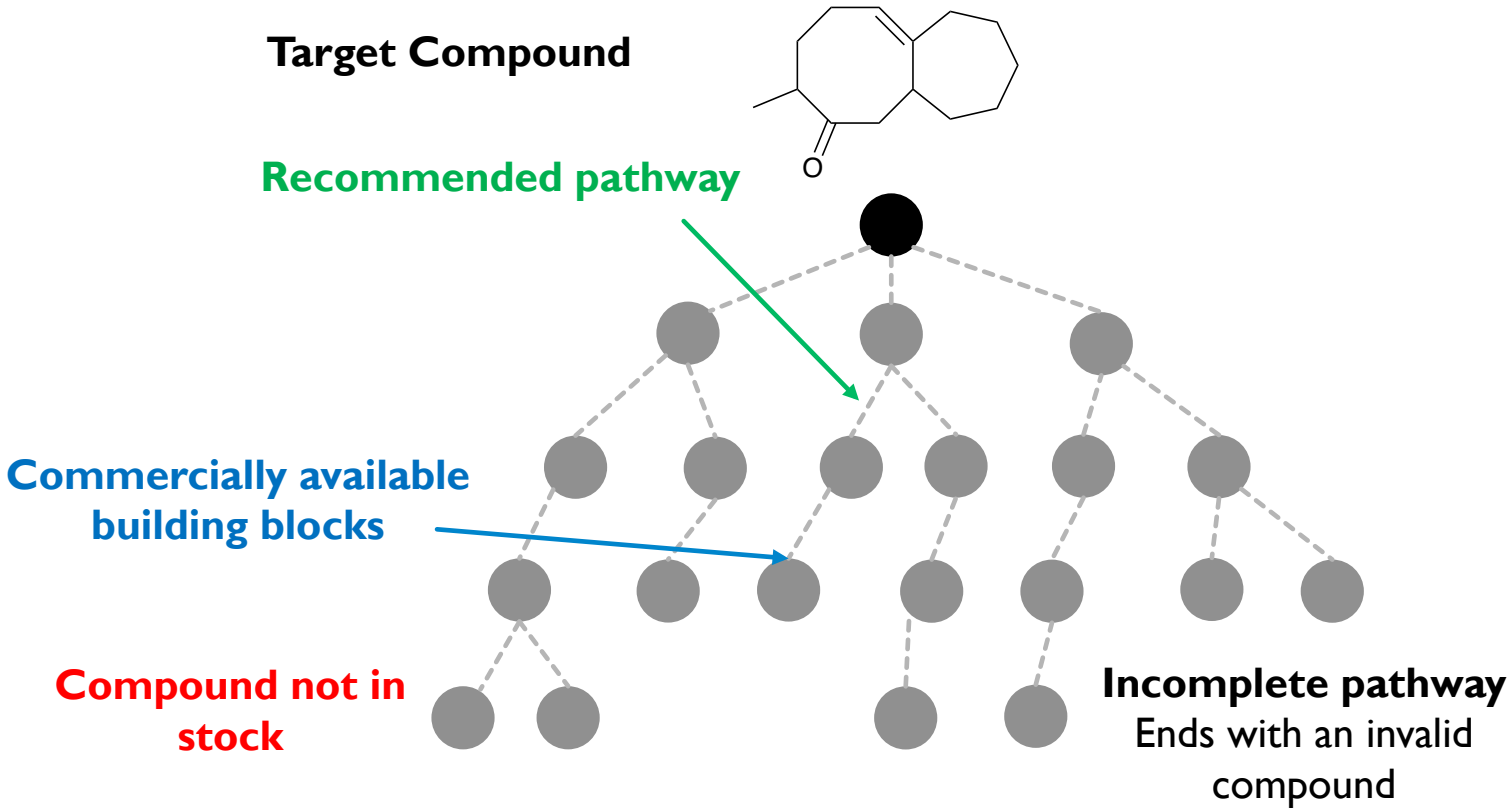


Is the template applicable in silico?

Product and template must have a substructure match

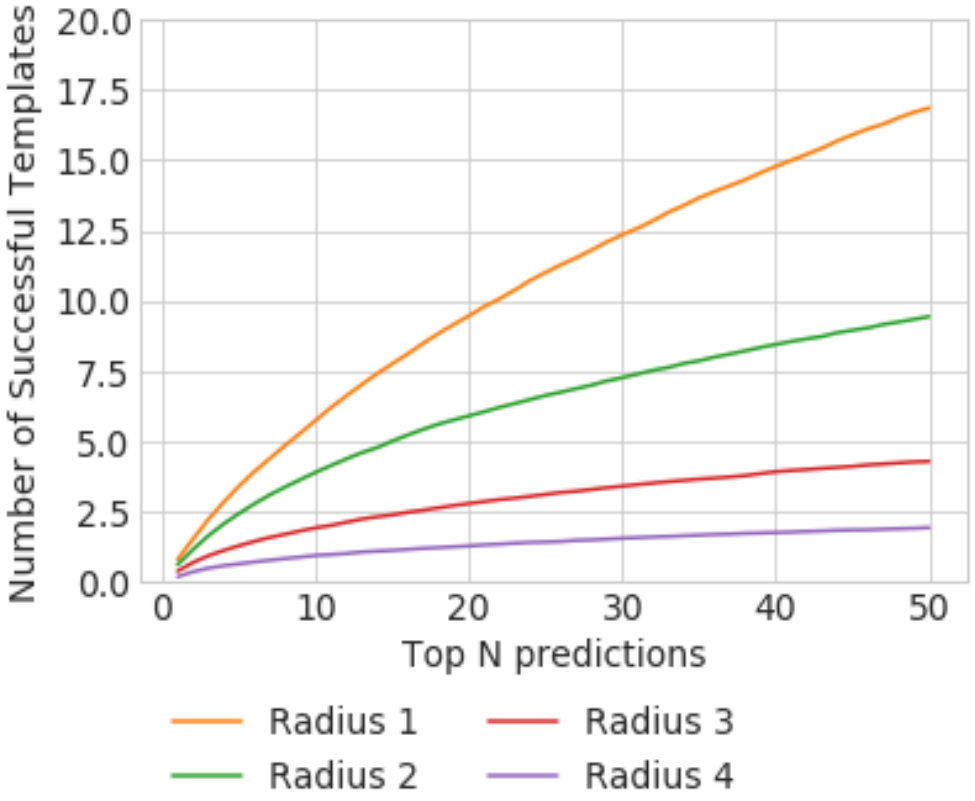
Thakkar et al., *Chemical Science*, **2020,** *11* (1), 154-168.
Segler et al., *Nature*, **2018**, 555, 604-610

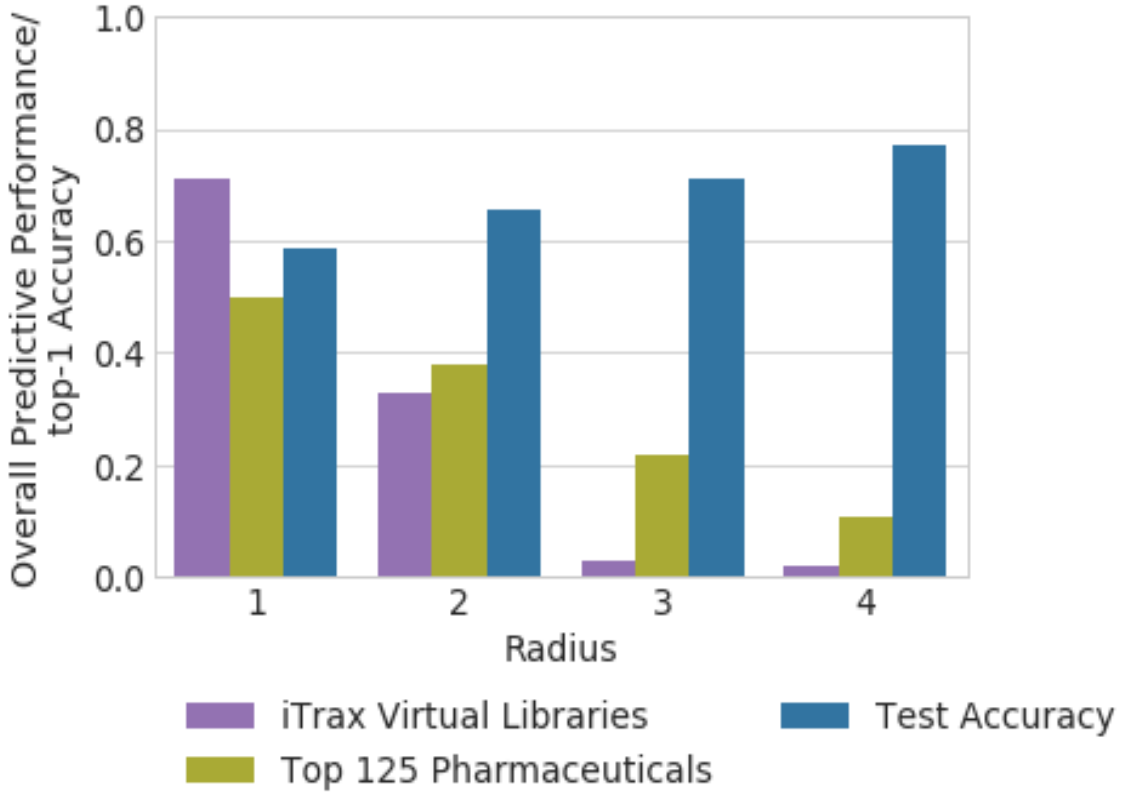
PREDICTING MULTISTEP PATHWAYS



Thakkar et al., *Chemical Science,* **2020,** *11* (1), 154-168.
Segler et al., *Nature*, **2018**, 555, 604-610

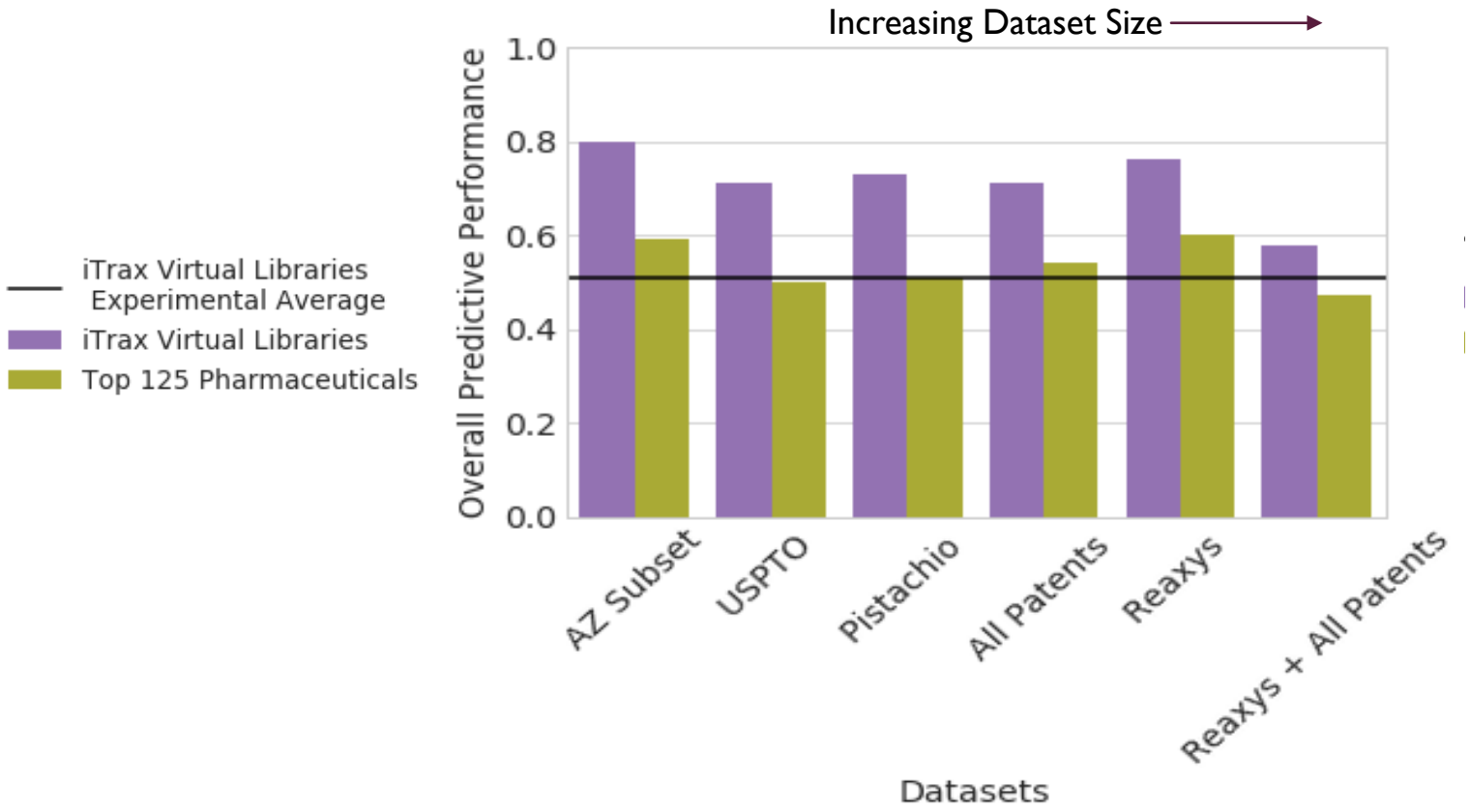
TEMPLATE SIZE AND PERFORMANCE





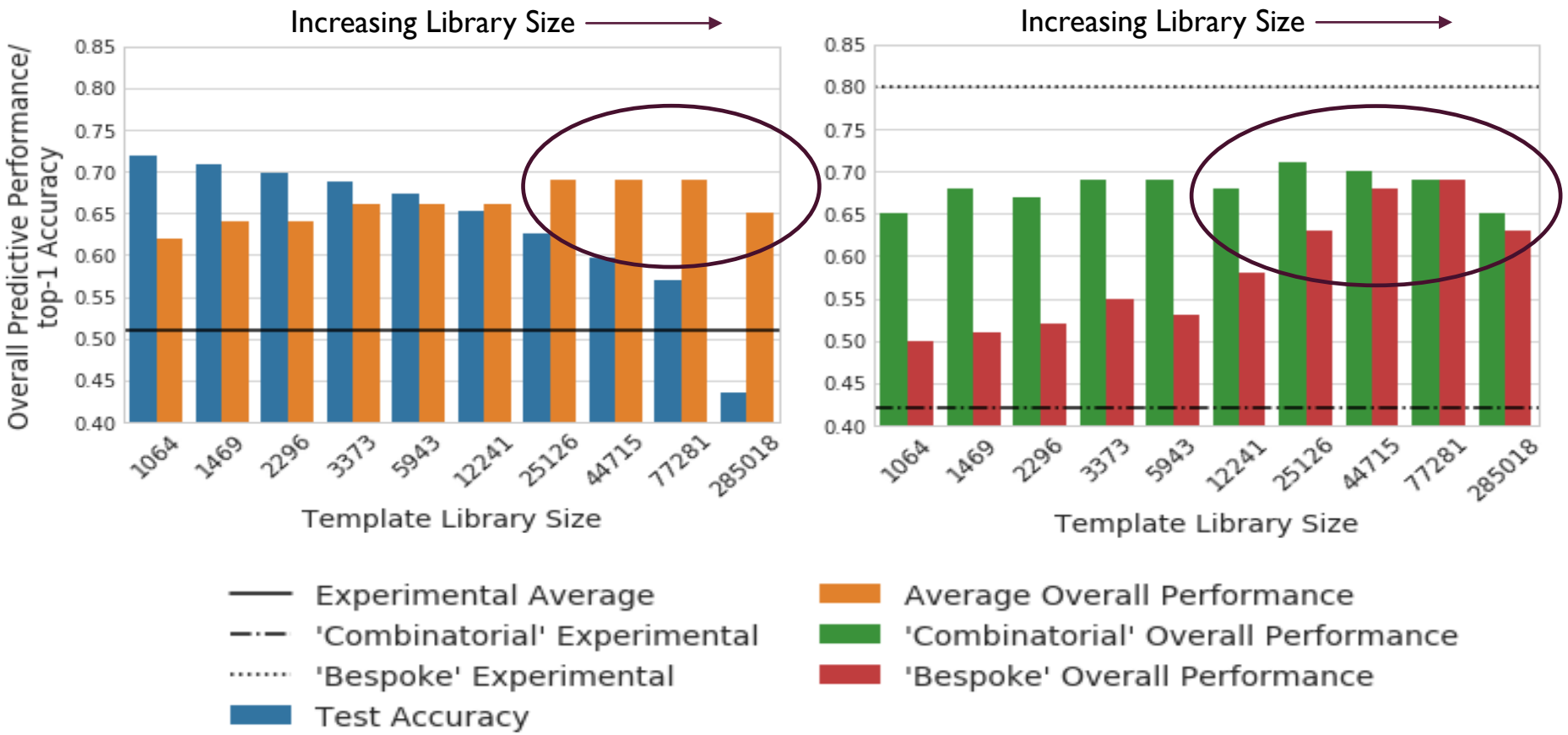
Thakkar et al., *Chemical Science*, **2020,** *11* (1), 154-168.

DATASETS AND THEIR PERFORMANCE



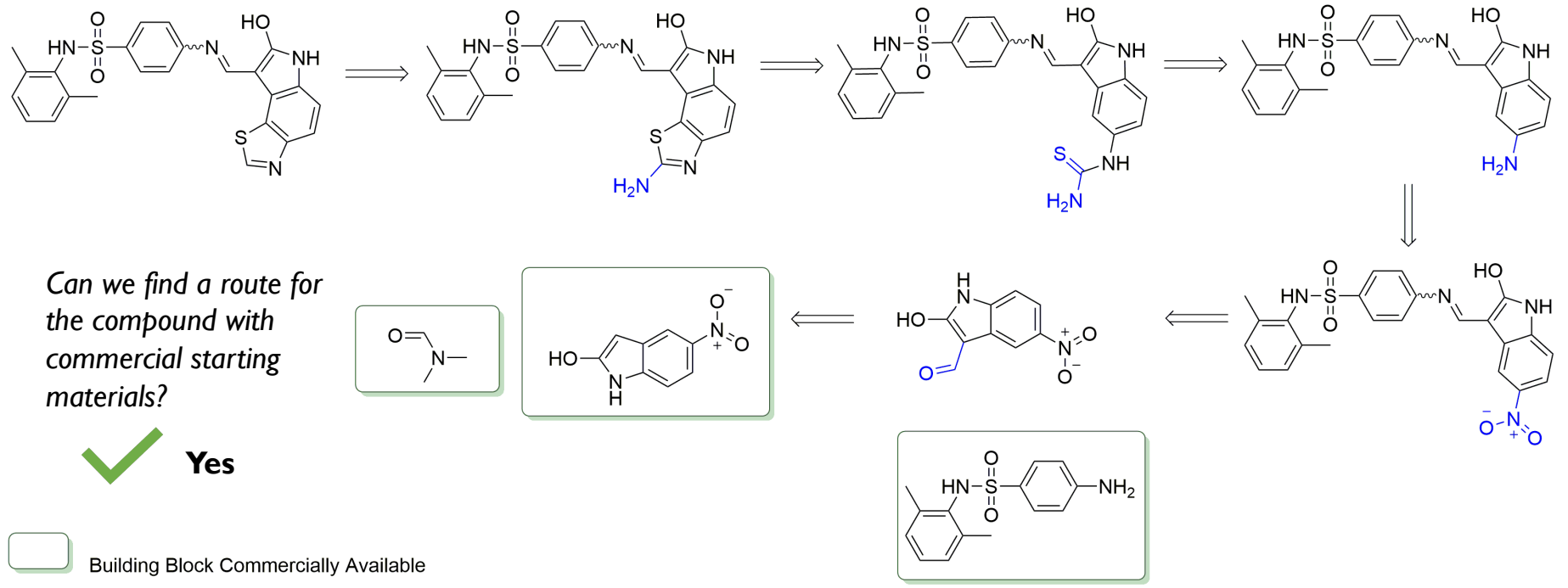
Thakkar et al., *Chemical Science*, **2020,** *11* (1), 154-168.

NUMBER OF TEMPLATES AND PERFORMANCE

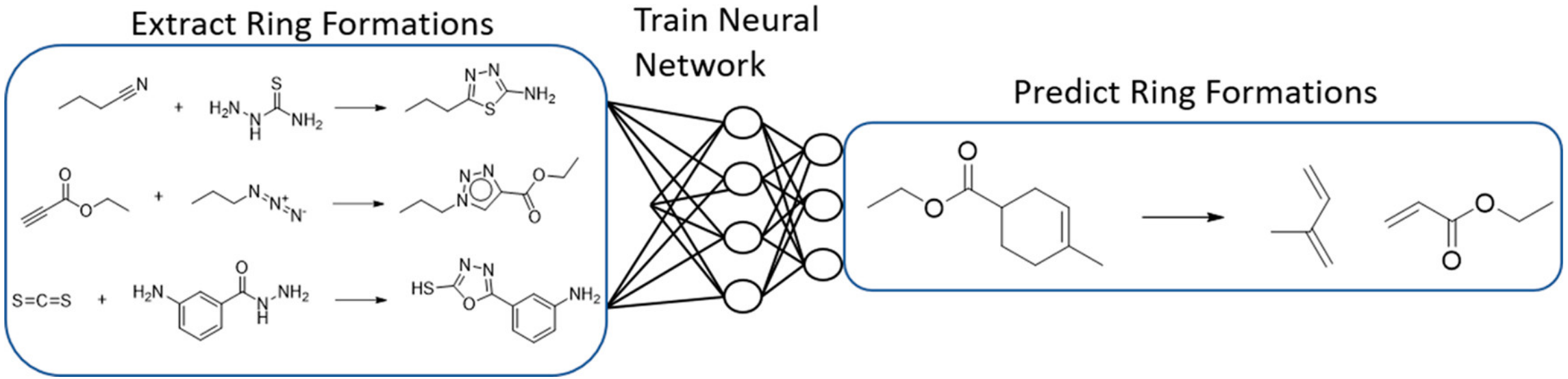


Thakkar et al., *Chemical Science*, **2020,** *11* (1), 154-168.

EXAMPLE AI GENERATED ROUTE



Thakkar et al., *Chemical Science*, **2020,** *11* (1)



RINGBREAKER

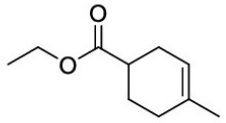
PREDICTION OF RING SYSTEMS

Thakkar et al. *Journal of Medicinal Chemistry,* **2020**, 63, 16, 8791-8808.

Highlights:
Artificial Intelligence in Drug Discovery special issue.

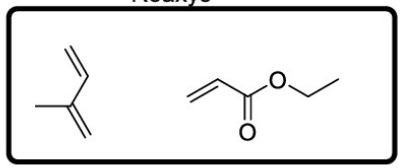
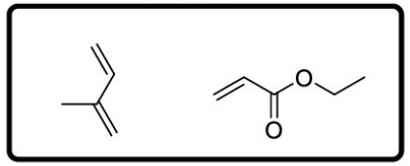
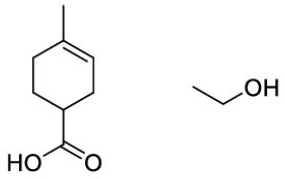
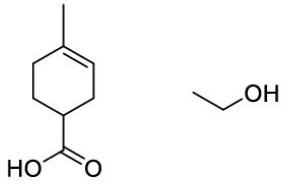
COMMON RING FORMING REACTIONS

Diels-Alder reaction

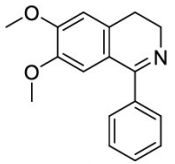


7

Org. Lett. 2006, 8, 12, 2487-2489

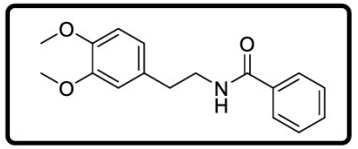
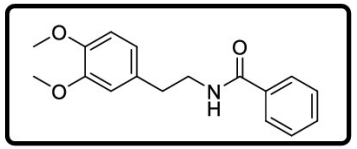
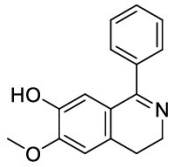
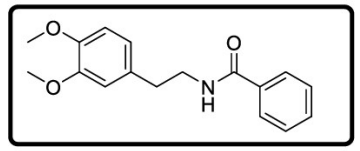
	Reaxys	USPTO
RingBreaker		
	rxid = 1405508 p(1) = 0.998 literature match = yes	id = US06703082B1 p(1) = 0.951 literature match = yes
Standard Model		
	rxid = 9644172 p(1) = 0.275 literature match = no	id = US05840917 p(1) = 0.383 literature match = no

Bischler-Napieralski reaction



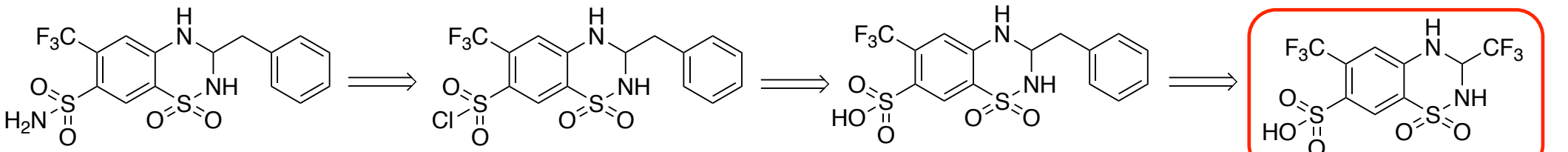
9

Org. Lett. 2008, 10, 16, 3485-3488

	Reaxys	USPTO
RingBreaker		
	rxid = 33663232 p(1) = 0.976 literature match = yes	id = US03966938 p(1) = 0.997 literature match = yes
Standard Model		
	rxid = 9149533 p(5) = 0.004 literature match = no	id = US03966938 p(1) = 0.966 literature match = yes

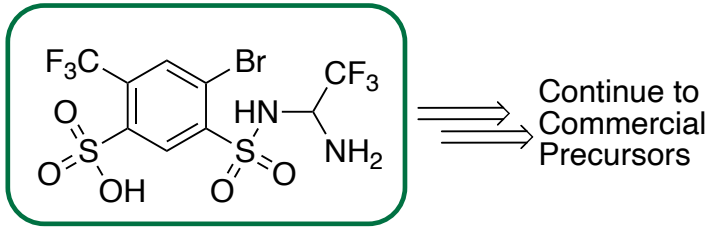
Thakkar et al. *Journal of Medicinal Chemistry,* **2020**, 63, 16, 8791-8808.

PRACTICAL USE CASE – INTERACTIVE MODE



- Standard Model STOPS
- Ring Disconnection NOT Predicted

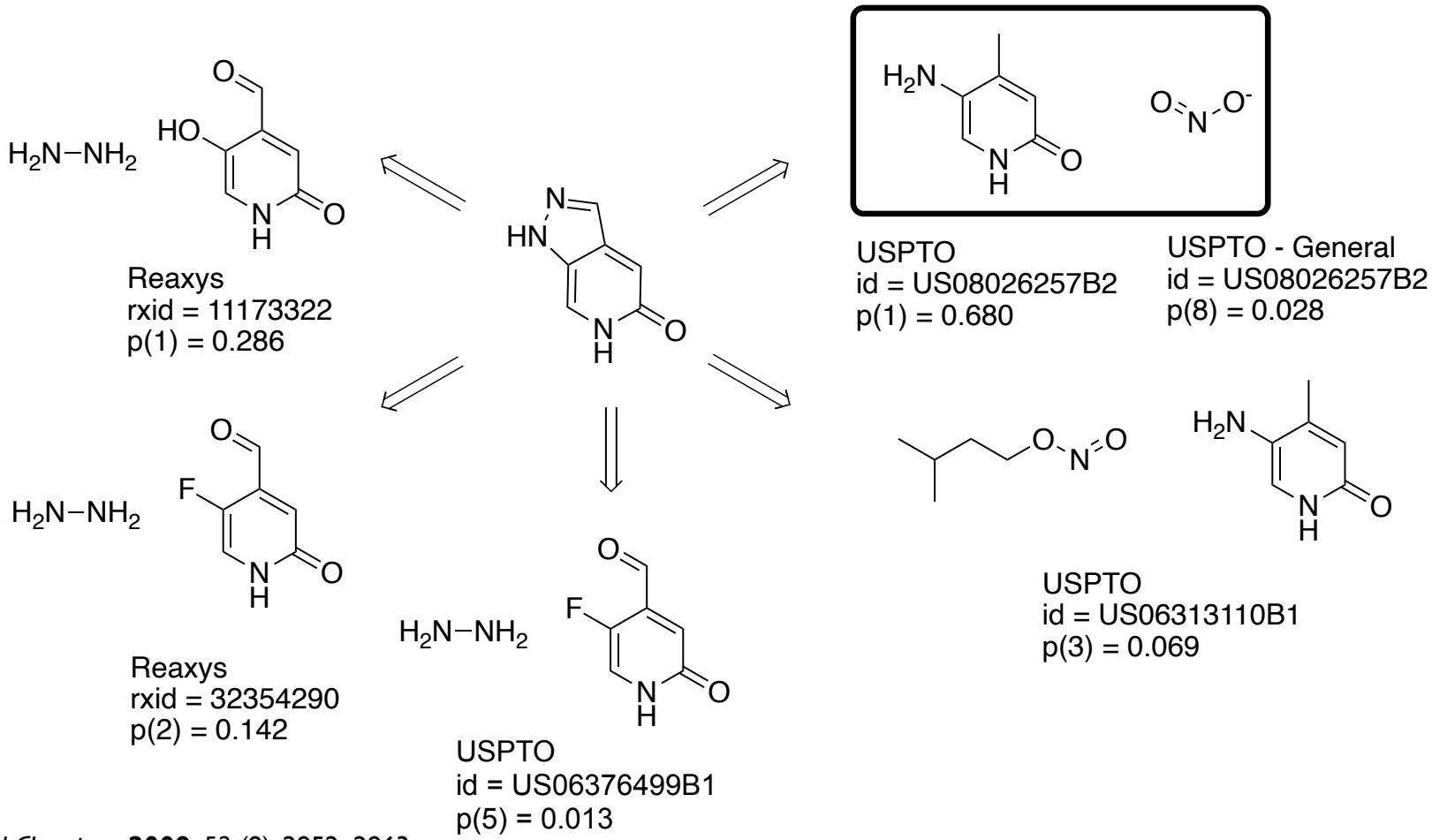
Use RINGBREAKER



Impacted ongoing project! Shortening route by 4 steps
(Not the route shown)

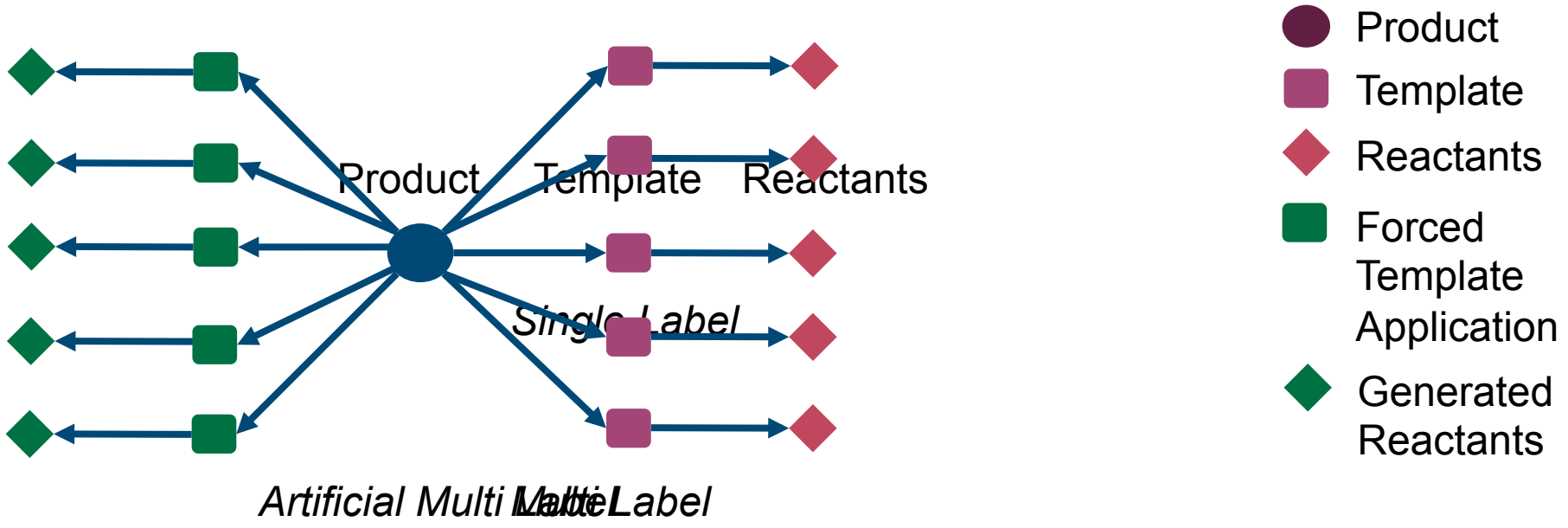
Thakkar et al. *Journal of Medicinal Chemistry,* **2020**, 63, 16, 8791-8808.

GENERALISING BEYOND THE TRAINING SET



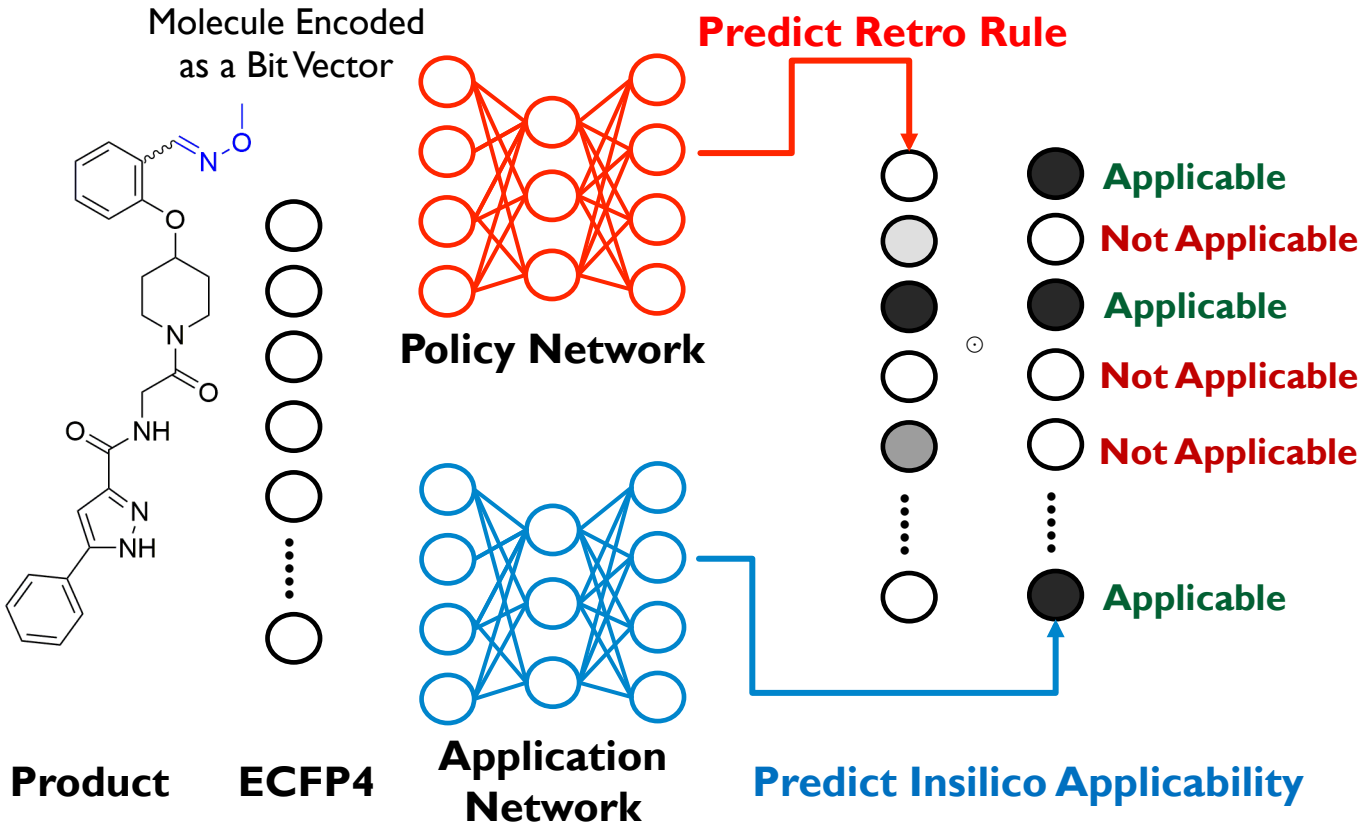
Pitt et al. *Journal of Medicinal Chemistry* **2009**, *52* (9), 2952–2963.
Silva Júnior et al. *RSC Advances* **2016**, *6* (27), 22777–22780.
Thakkar et al. *Journal of Medicinal Chemistry,* **2020**, 63, 16, 8791-8808.

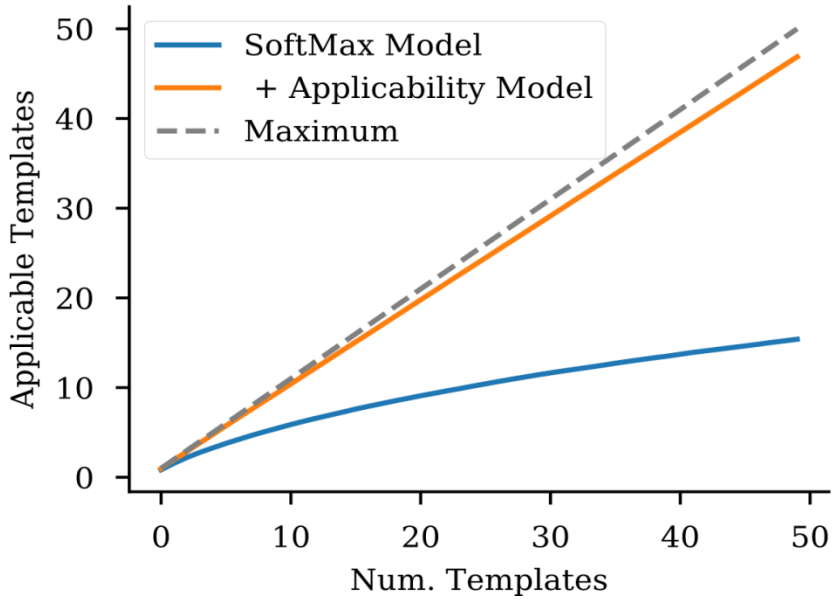
REACTION DATA AUGMENTATION – ARTIFICIAL LABELS



Thakkar et al. *Journal of Medicinal Chemistry,* **2020**, 63, 16, 8791-8808.
Bjerrum E, **Thakkar A** et al. MLST, **2020**, 2

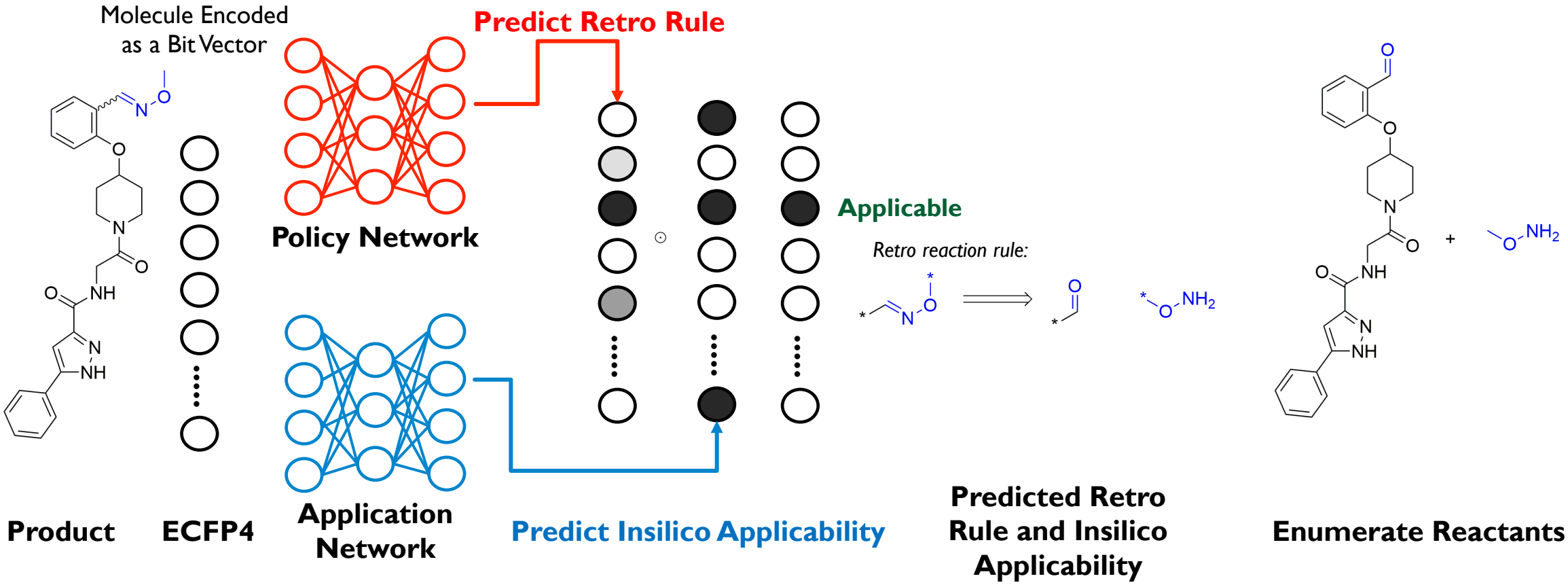
IMPROVING TEMPLATE PRIORITISATION



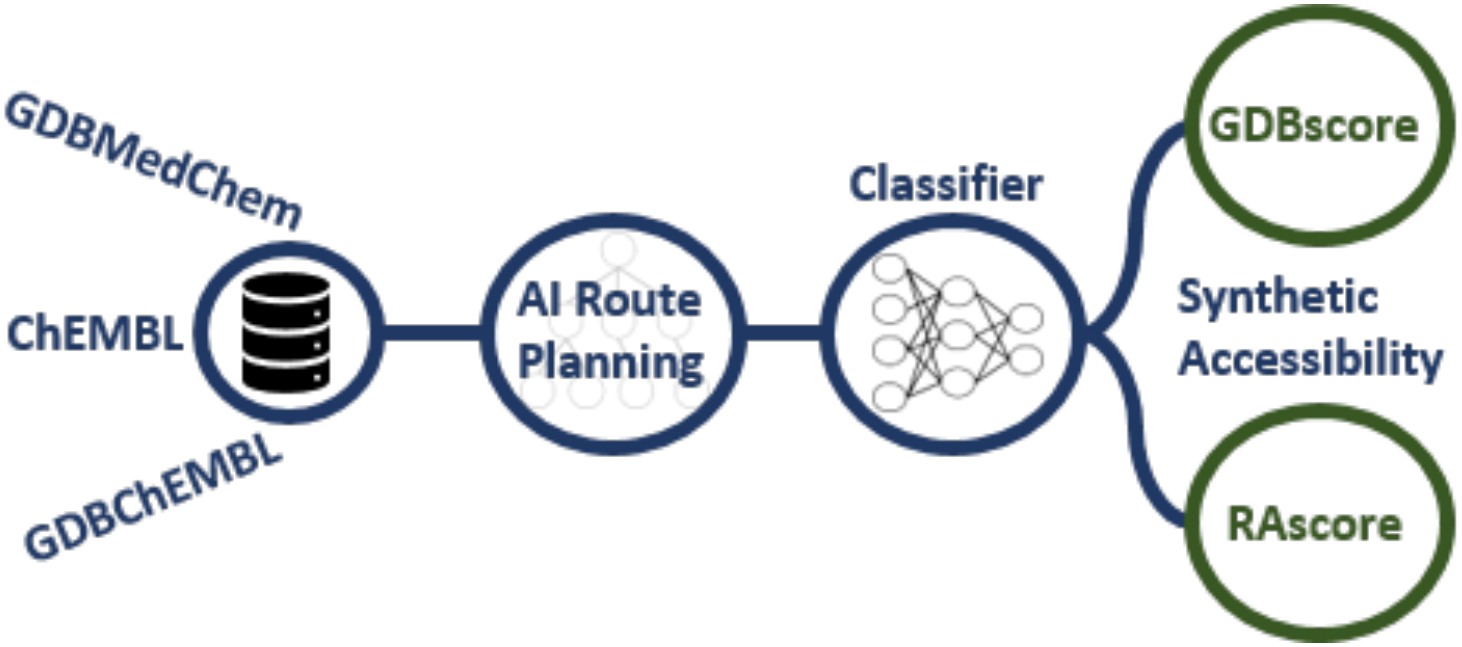


Thakkar et al., *Chemical Science*, **2020,** *11* (1), 154-168.
Bjerrum E, **Thakkar A** et al. MLST, **2020**, 2

IMPROVING TEMPLATE PRIORITISATION



Thakkar et al., *Chemical Science*, **2020,** *11* (1), 154-168.
Bjerrum E, **Thakkar A** et al. MLST, **2020**, 2

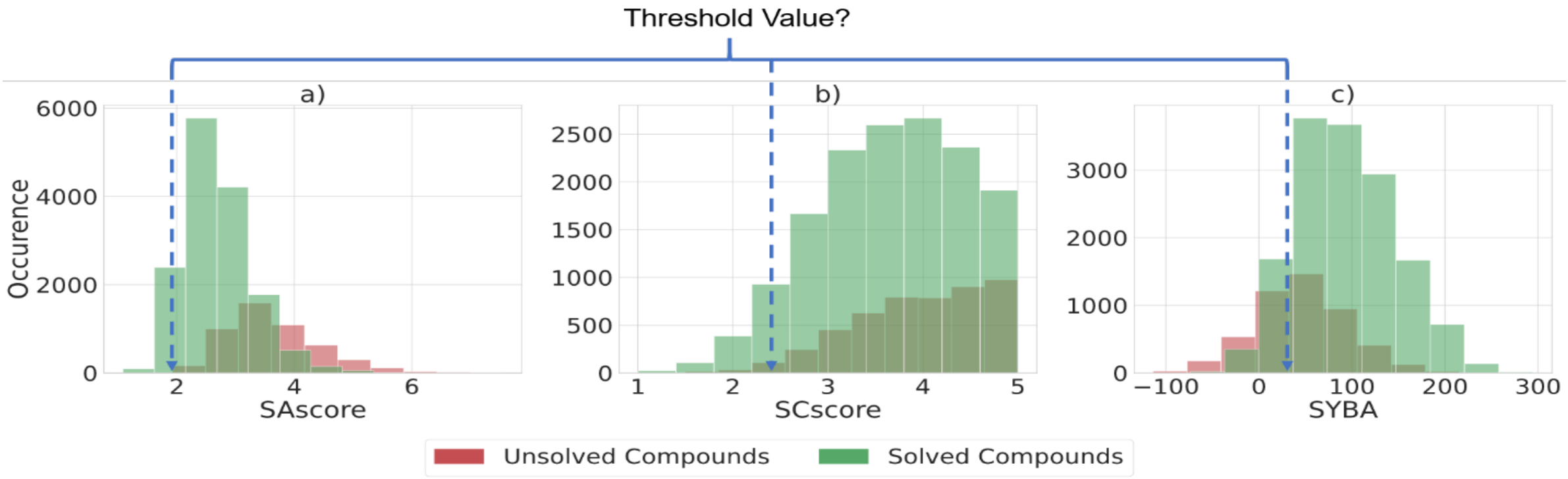


RETROSYNTHETIC ACCESSIBILITY SCORE (RASCORE)

Thakkar et al. *Chem. Sci.* **2021**, *12* (9), 3339–3349

Highlights:
Most popular 2021 physical and theoretical chemistry articles and Editor's Choice – Graeme Day

EXISTING SYNTHETIC COMPLEXITY SCORES

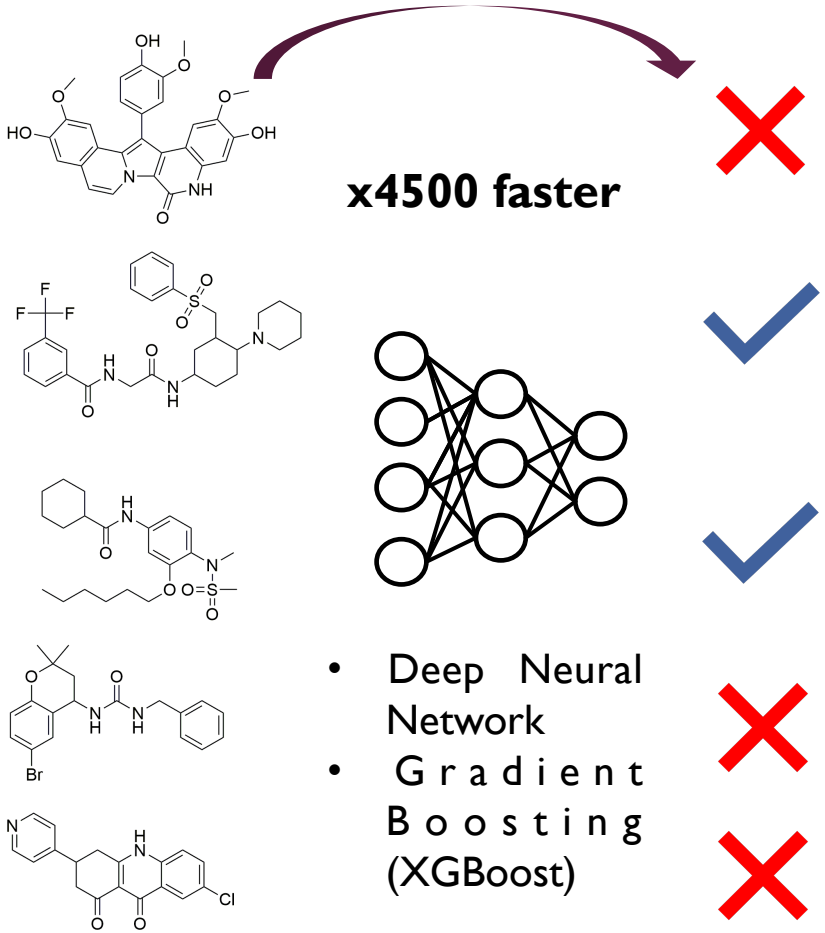


Ertl et al., *J. Cheminf.*, 2009, **1**(1), 8
Coley et al., *J. Chem. Inf. Model.*, 2018, **58**(2), 252–261
Voršilák et al., *J. Cheminf.*, 2020, **12**(1), 35
Thakkar et al. *Chem. Sci.* **2021**, *12* (9), 3339–3349

RETROSYNTHETIC ACCESSIBILITY SCORE (RASCORE)

239 days

200,000 compounds from ChEMBL



77 minutes

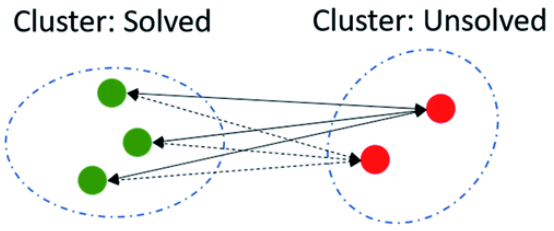
Machine Learning Model Based on Predictions

Thakkar et al. *Chem. Sci.* **2021**, *12* (9), 3339–3349

[a] Expressed in days taken on a single machine with 8 CPUs and 64 GB of RAM (no GPU required), rounded to the nearest day. The time taken in minutes for the neural network classifier with ECFP6 counted fingerprints is also given for comparative purposes. The neural network classifier, RAscore, is able to reproduce the results obtained from AiZynthFinder in a fraction of the time taken to predict full retrosynthetic routes. [b]Rascore [c]GDBscore

MODEL ASSESSMENT

	ChEMBL	GDBChEMBL	GDBMedChem
Percentage Solved	75.21	25.54	20.79
Size	200,000	100,000	100,000
AiZynthFinder Run Time (days)	239	149	151
Score Run Time (mins)	79 [b)]	30 [c)]	30 [c)]



$$average\ linkage = \left| \frac{1}{n_{solved} n_{unsolved}} \sum_{i=1}^{n_{solved}} \sum_{j=1}^{n_{unsolved}} D\left(x_{solved,i}, x_{unsolved,j}\right) \right|$$

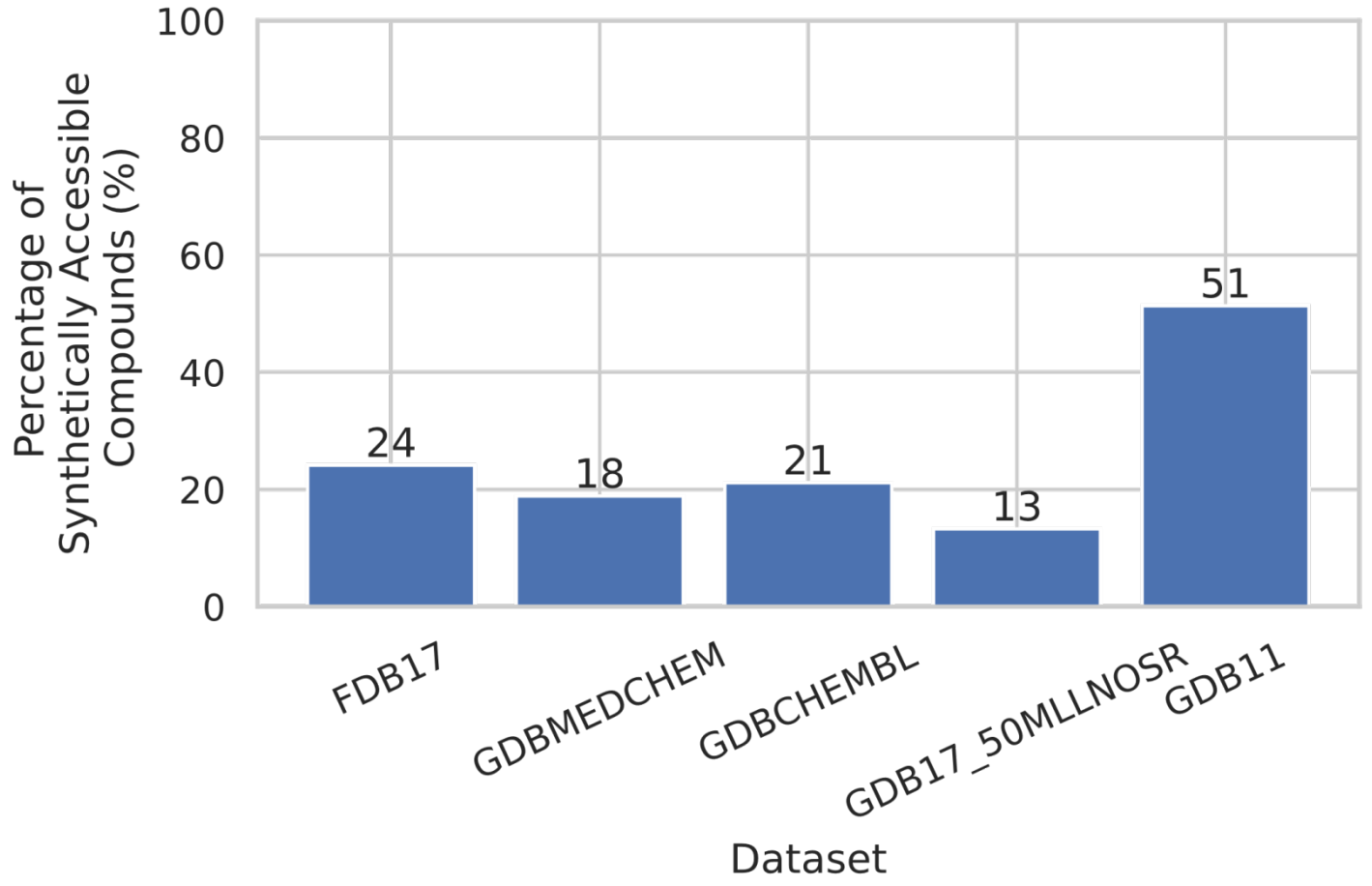
Average distance between all pairs of items

Model	AUC	Accuracy	Precision	Recall	Average Linkage
N N (RAscore)	0.93	0.90	0.92	0.95	0.69
XGB	0.95	0.91	0.92	0.96	0.65

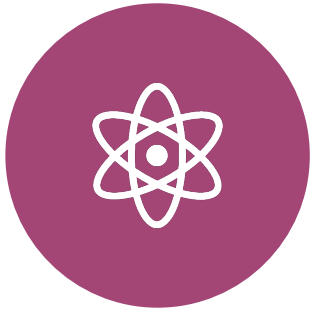
Thakkar et al. *Chem. Sci.* **2021**, *12* (9), 3339–3349

[a)] Expressed in days taken on a single machine with 8 CPUs and 64 GB of RAM (no GPU required), rounded to the nearest day. The time taken in minutes for the neural network classifier with ECFP6 counted fingerprints is also given for comparative purposes. The neural network classifier, RAscore, is able to reproduce the results obtained from AiZynthFinder in a fraction of the time taken to predict full retrosynthetic routes. [b)]Rascore [c)]GDBscore

ASSESSING THE GDB CHEMICAL SPACE



Curently individual components and we want to combine them – lower barrier to entry



GDB CHEMICAL SPACE

AI PLANNED SYNTHESIS

BROWSER BASED EXPLORATION

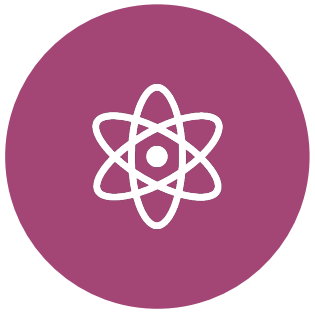
EXPERIMENTAL ENGAGEMENT

ACCESSING GDB CHEMICAL SPACE

ENABLING EXPERIMENTATION BY LINKING CHEMICAL LIBRARY VISUALISATION

Unpublished – Corresponding Slides Removed

LINKING LIBRARY VIZUALISATION TO SYNTHESIS



GDB CHEMICAL SPACE

AI PLANNED SYNTHESIS

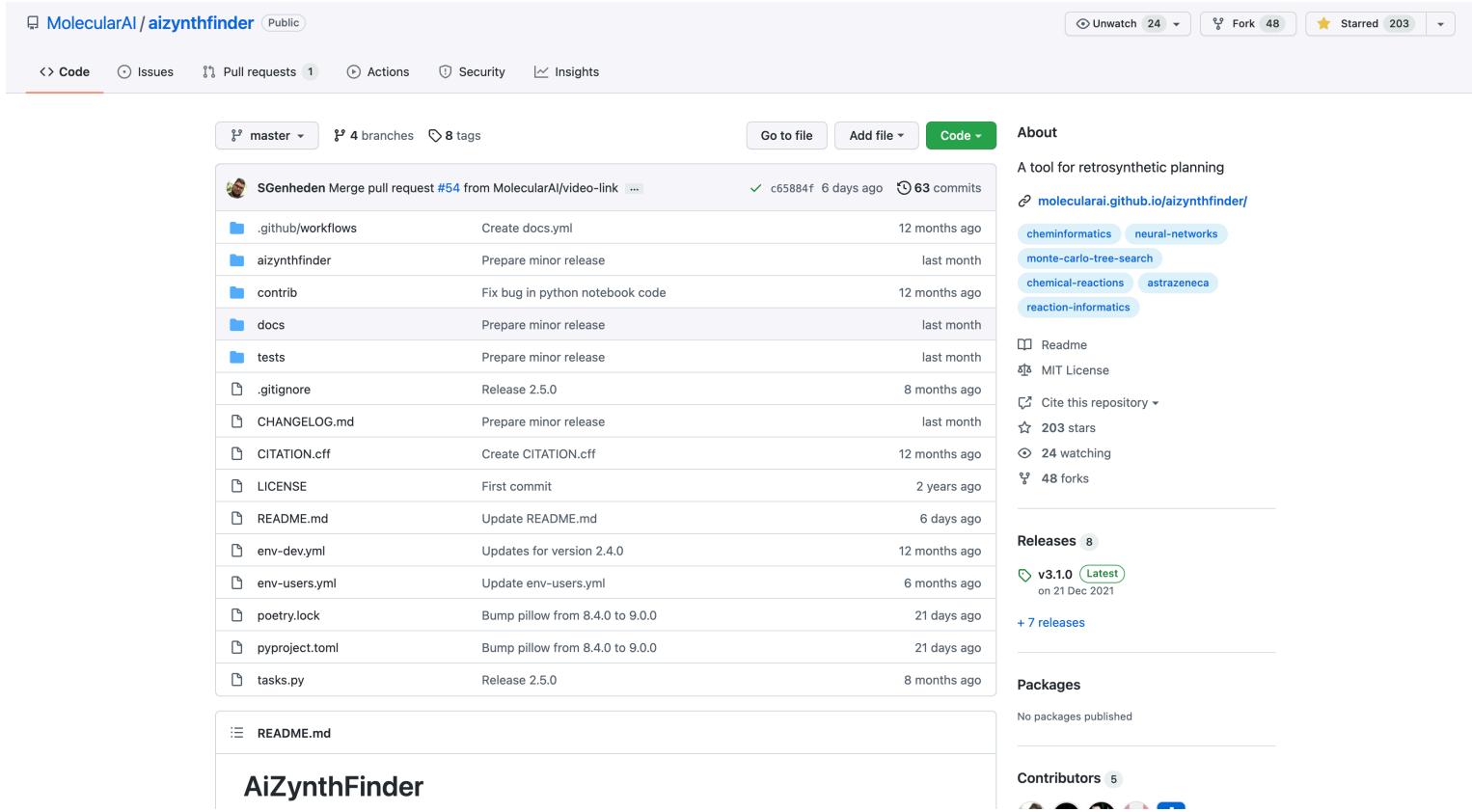
BROWSER BASED EXPLORATION

EXPERIMENTAL ENGAGEMENT

- × Scripting or command line knowledge
- × Step wise compound submission
- × Wait for batch runs
- × Switching between tools for visualisation, calculation of properties, synthesis prediction, prioritisation

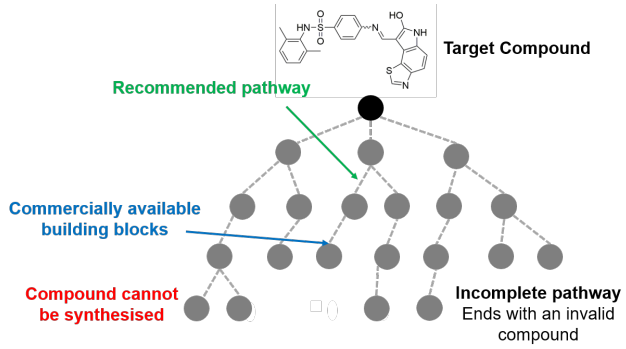
- ✓ One tool
- ✓ Precomputed routes – fast access
- ✓ Allows whole library to be visualised
- ✓ Enables easier prioritisation using synthetic route information

SUMMARY

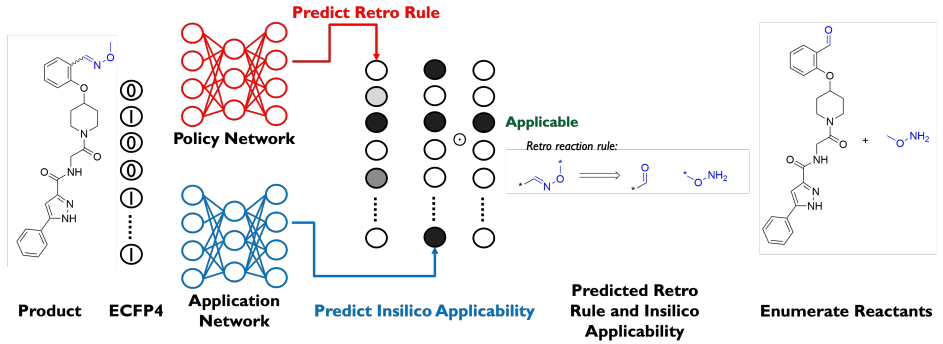

MolecularAI / aizynthfinder
Public
Unwatch 24
Fork 48
Starred 203
Code
Issues
Pull requests 1
Actions
Security
Insights
master
4 branches
8 tags
Go to file
Add file
Code
SGenheden Merge pull request #54 from MolecularAI/video-link
c65884f 6 days ago
63 commits
.github/workflows Create docs.yml 12 months ago
aizynthfinder Prepare minor release last month
contrib Fix bug in python notebook code 12 months ago
docs Prepare minor release last month
tests Prepare minor release last month
.gitignore Release 2.5.0 8 months ago
CHANGELOG.md Prepare minor release last month
CITATION.cff Create CITATION.cff 12 months ago
LICENSE First commit 2 years ago
README.md Update README.md 6 days ago
env-dev.yml Updates for version 2.4.0 12 months ago
env-users.yml Update env-users.yml 6 months ago
poetry.lock Bump pillow from 8.4.0 to 9.0.0 21 days ago
pyproject.toml Bump pillow from 8.4.0 to 9.0.0 21 days ago
tasks.py Release 2.5.0 8 months ago
README.md
AiZynthFinder
About
A tool for retrosynthetic planning
molecularai.github.io/aizynthfinder/
cheminformatics
neural-networks
monte-carlo-tree-search
chemical-reactions
astrazeneca
reaction-informatics
Readme
MIT License
Cite this repository
203 stars
24 watching
48 forks
Releases 8
v3.1.0 Latest
on 21 Dec 2021
+ 7 releases
Packages
No packages published
Contributors 5

SUMMARY

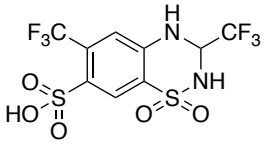
AiZynthFinder – Open-Source Retrosynthetic Planning



Data Augmentation – Artificial Labels

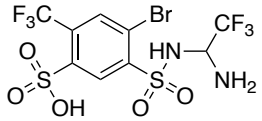


RingBreaker

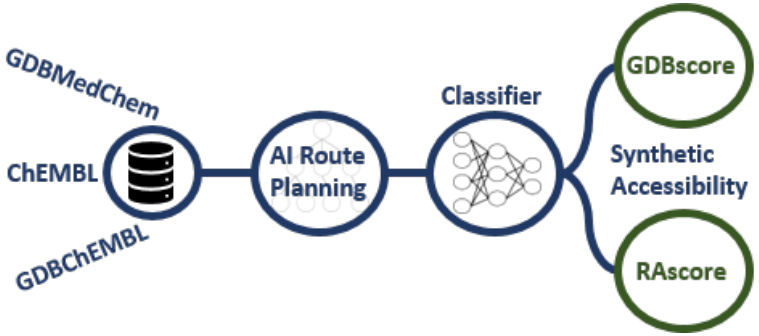


- Standard Model STOPS
- Ring Disconnection NOT Predicted

Use RINGBREAKER



Retrosynthetic Accessibility



GDBRouteBrowser

ACKNOWLEDGEMENTS

Professor Jean-Louis Reymond
Dr Ola Engkvist (AZ)
Dr Esben Jannik Bjerrum (AZ)
Dr Samuel Genheden (AZ)

Reymond Group, University of Bern, Switzerland
Molecular AI, AstraZeneca, Sweden
Global Chemistry Community, AstraZeneca
Elsevier
NextMove Software

The reviewers for their useful feedback.

amol.thakkar@unibe.ch