UNIVERSITÄT BERN

10th RDKit UGM

14-15 October 2021

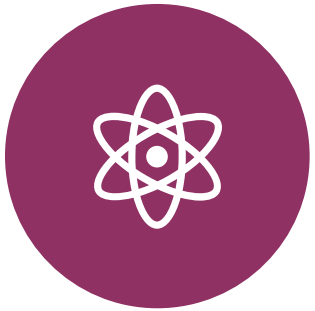
15th October 2021

BROWSER BASED EXPLORATION OF THE GDB CHEMICAL SPACE USING AI PLANNED SYNTHESIS FACILITATES EXPERIMENTAL ENGAGEMENT

AMOL THAKKAR

INTRODUCTION - OVERVIEW

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



GDB CHEMICAL SPACE

AI PLANNED SYNTHESIS

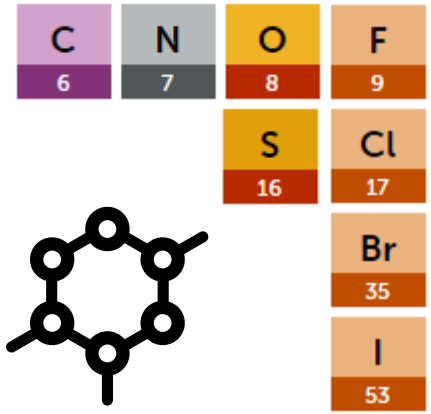
BROWSER BASED EXPLORATION

EXPERIMENTAL ENGAGEMENT

Enabling experimentation by linking chemical library visualisation to predicted synthesis in one tool

GDB CHEMICAL SPACE – HOW MANY MOLECULES ARE POSSIBLE?

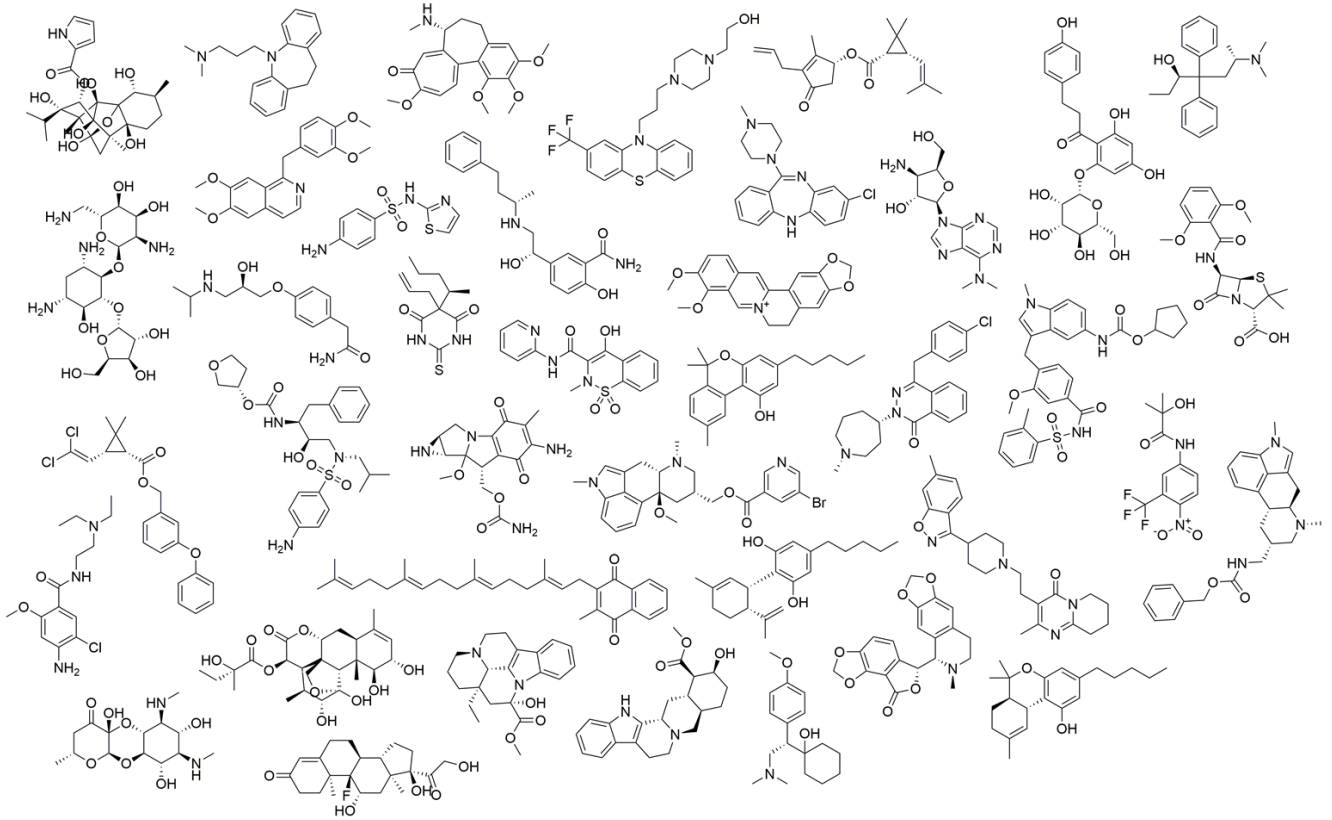
Target Identification → **Enumeration** → Filtering → Visualisation → Synthesis Planning → Wet Lab Candidates



ca. 166 billion

GDB17

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



ca. 10 million

FDB17

Visini et. al., *J. Chem. Inf. Model.* **2017**, 57, 4, 700–709

ca. 10 million

GDBMedChem

Awale et. al., *Mol. Inf.* **2019**, *38*, 1900031

ca. 10 million

GDBChEMBL

Bühlmann et. al., *Front. Chem.* **2020**, *8* (46)

TACKLING COMPUTER AIDED SYNTHESIS COLLABORATIVELY

SYNTHESIS PLANNING

WET LAB CANDIDATES

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

UNIVERSITÄT BERN

MARIE CURIE ACTIONS

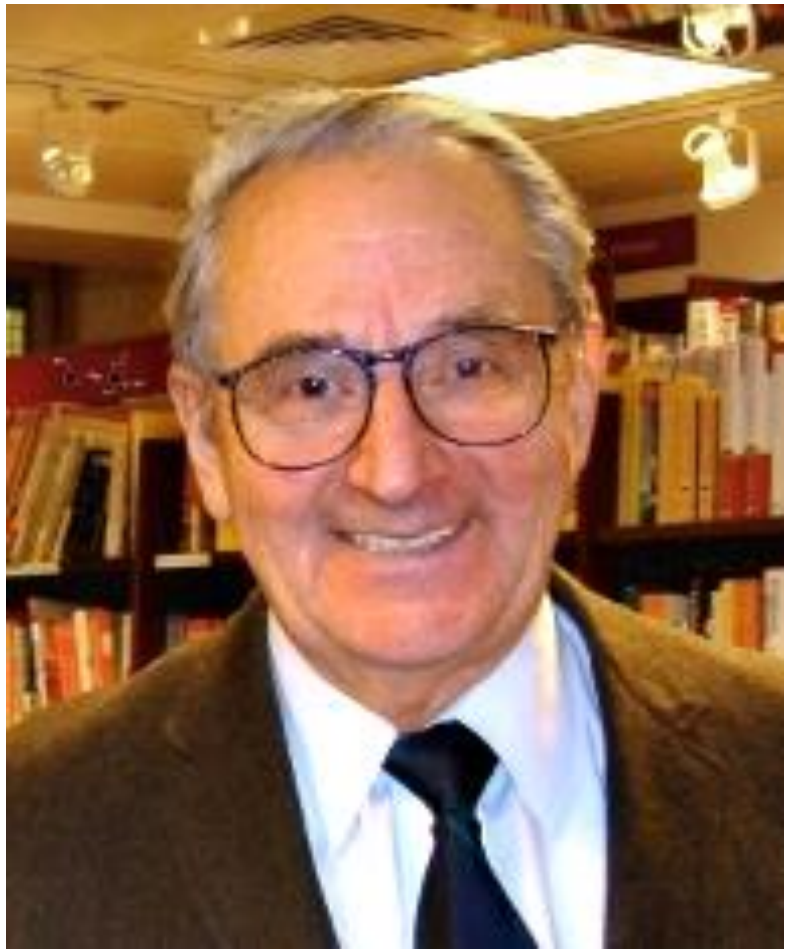
AstraZeneca

PRIOR ART

- Guide chemists in the synthesis of molecules
- Codify and organise the techniques used in organic synthesis
- Propose a variety of synthetic routes given any molecule

The responsibility for the final evaluation of the merit of the routes lies with the chemist.

We need a method for getting feedback on lots of routes quickly – improving synthesis plans by iteration



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 organiz 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

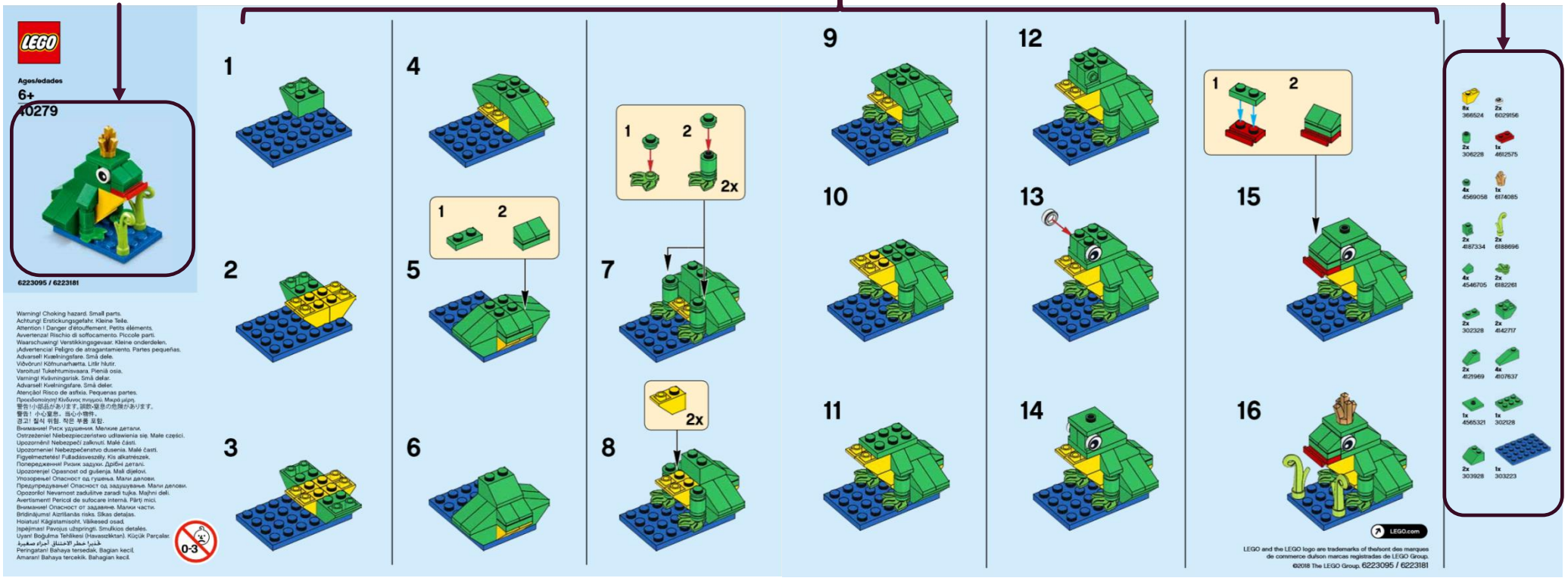
LEGO – ANALOGY TO SYNTHESIS PLANNING

Building Lego as an analogy to synthetic route planning

Target Molecule

Sequence of steps to build target

Known Building Blocks

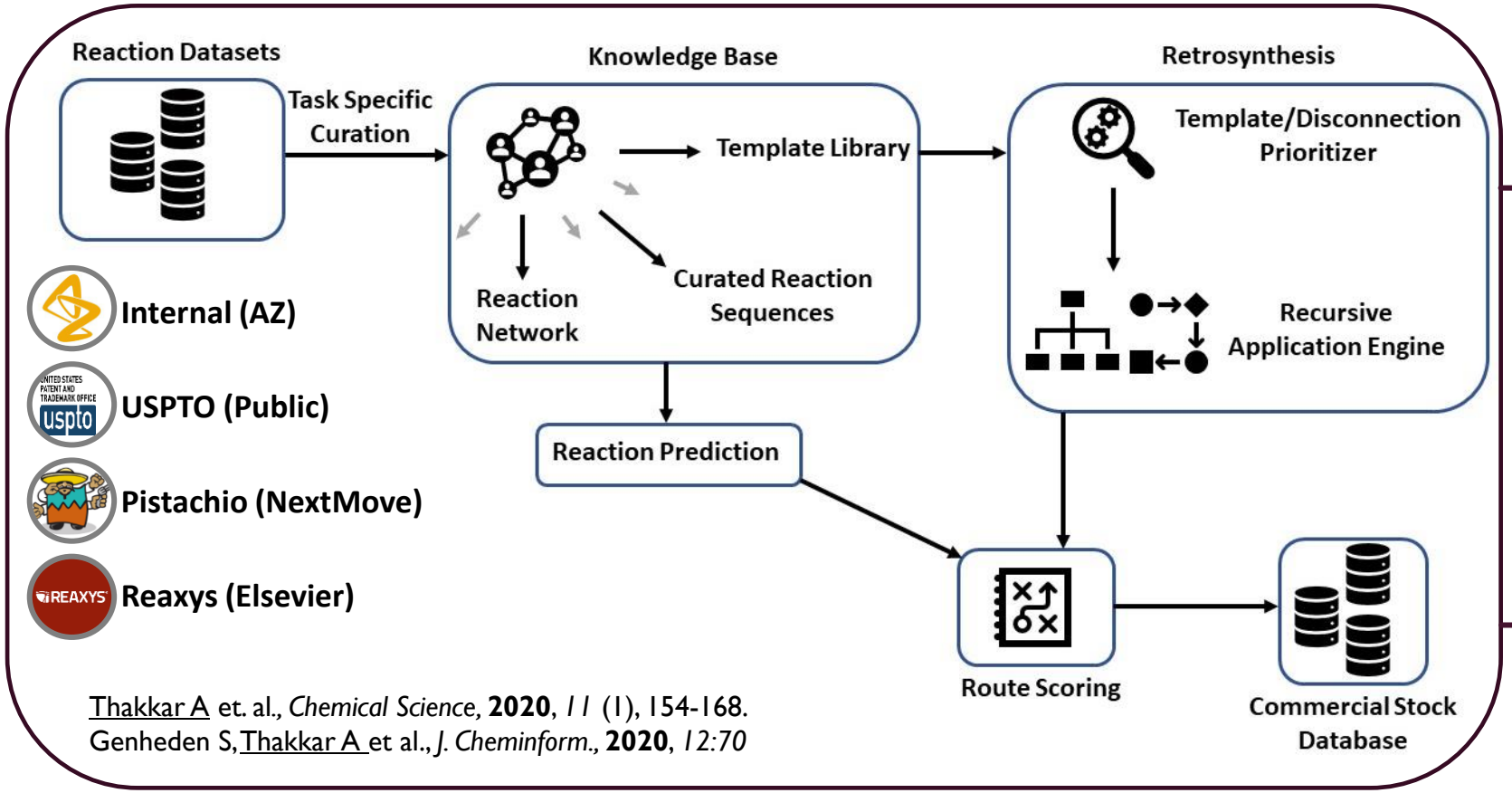


OPEN-SOURCE APPROACHES FOR SYNTHESIS PLANNING

Code

Paper

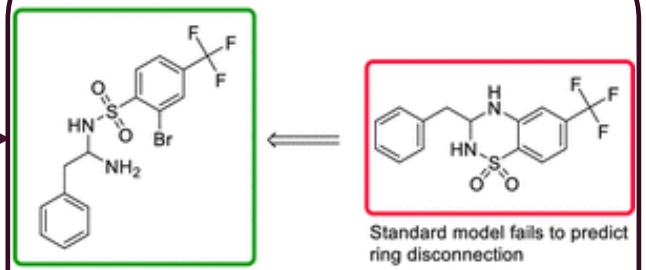
AiZynthFinder – Retrosynthetic Planning



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

RingBreaker

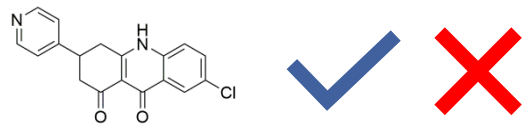
Adaptation for ring systems



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

RA score

Rapid retrosynthetic feasbility assesment



Thakkar A et. al., . *Chemical Science*, **2021**, Advance Article

TEMPLATE EXTRACTION

Code

Paper

Dataset

Proprietary

Internal (AZ)

Literature

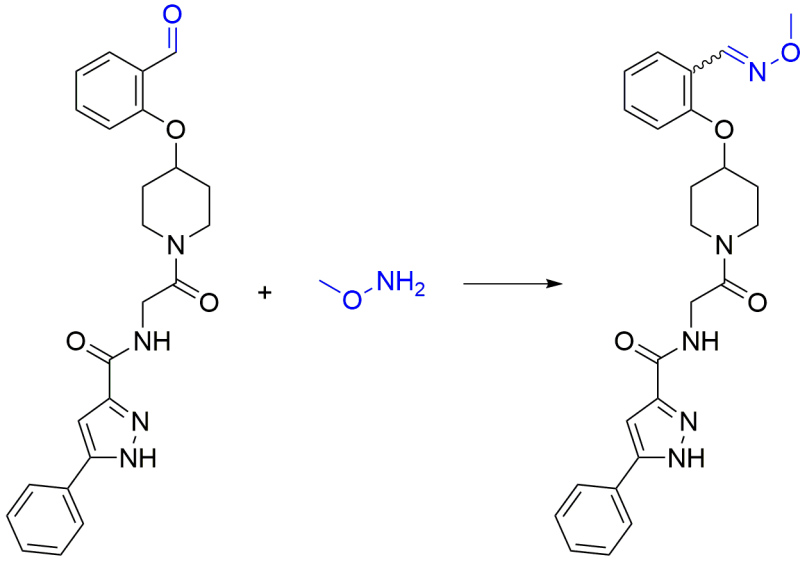
Reaxys (Elsevier)

Patents

USPTO (Public)

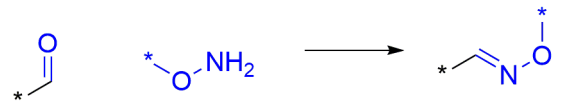
Pistachio (NextMove)

Reaction in Dataset

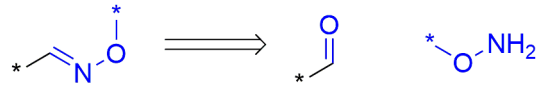


Extract Rules

Forward reaction rule:



Retro reaction rule:

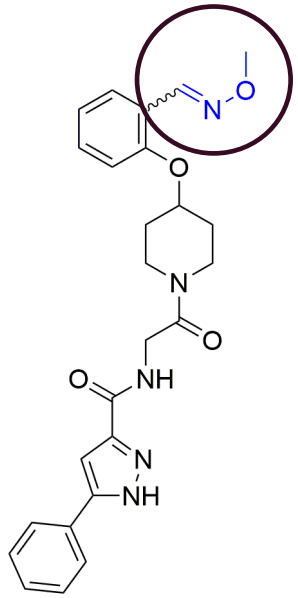


Thakkar et al., *Chemical Science*, **2020,** *11* (1), 154-168.
Segler et al., *Nature*, **2018**, 555, 604-610

NEURAL NETWORK TRAINING – TEMPLATE PRIORITISATION

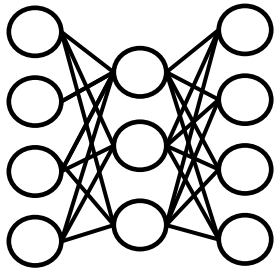
Code

Paper



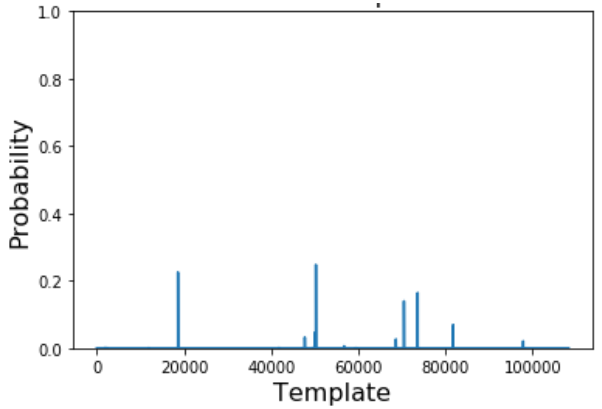
Product

ECFP4

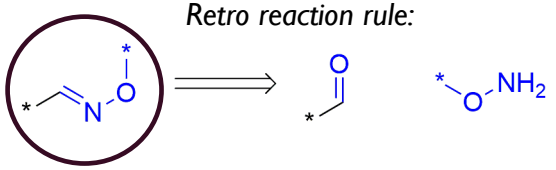


Priotisation Network

Predict Retrosynthesis Template



Retro reaction rule:



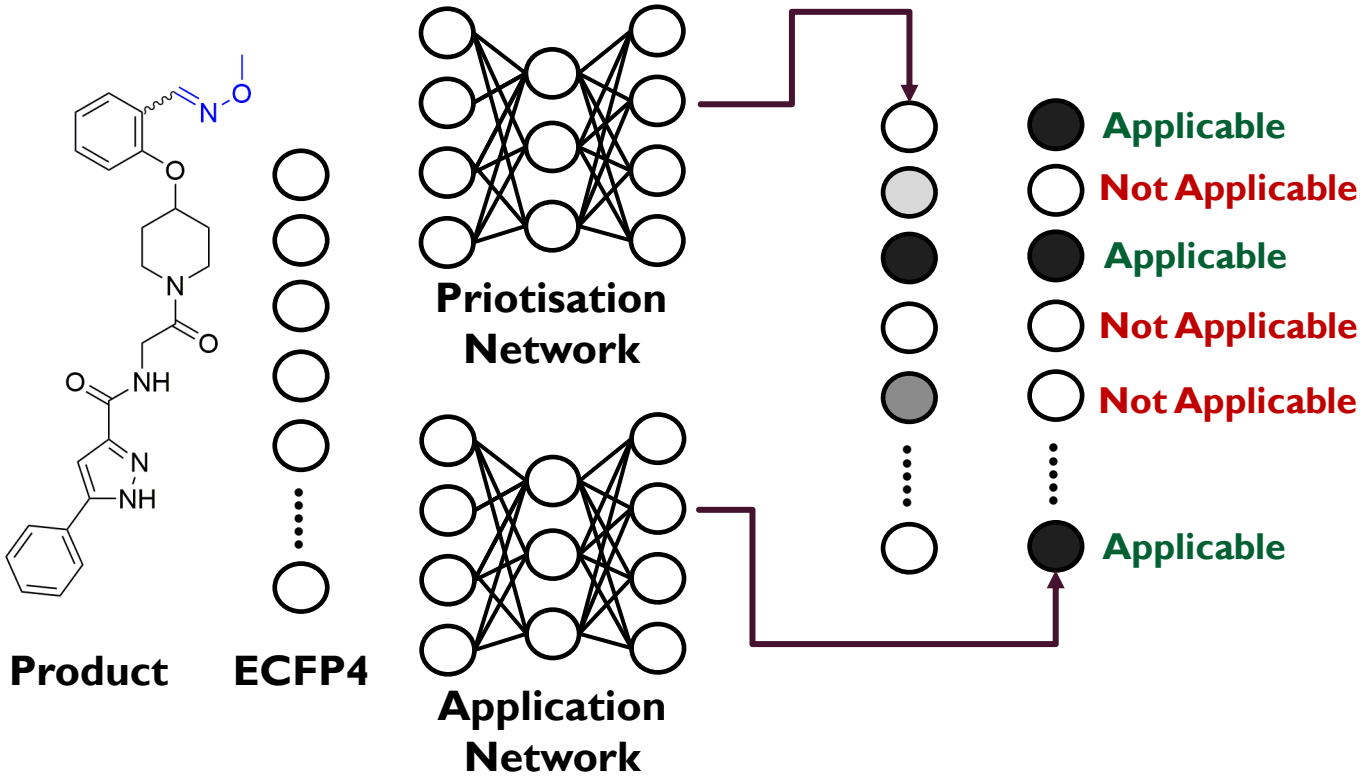
Is the template applicable in silico?

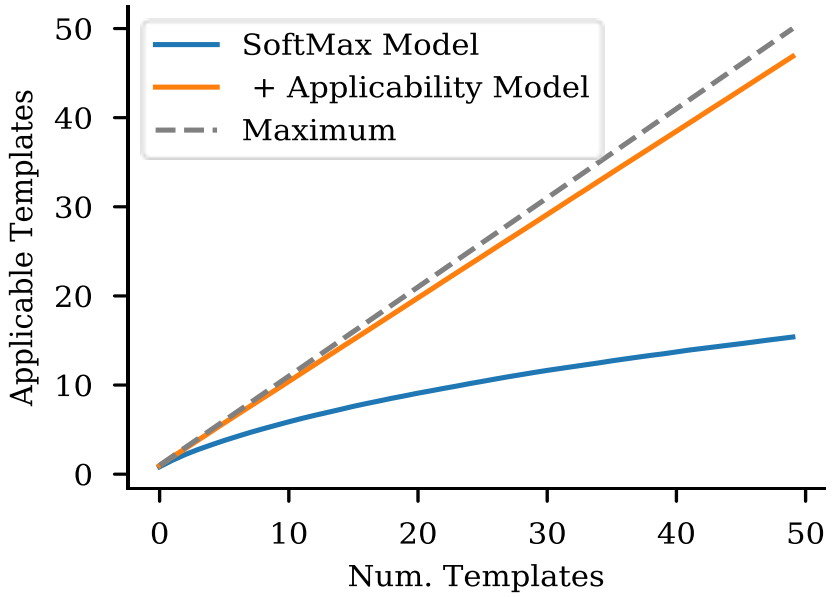
Product and template must have a substructure match

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

IMPROVING TEMPLATE PRIORITISATION

Paper

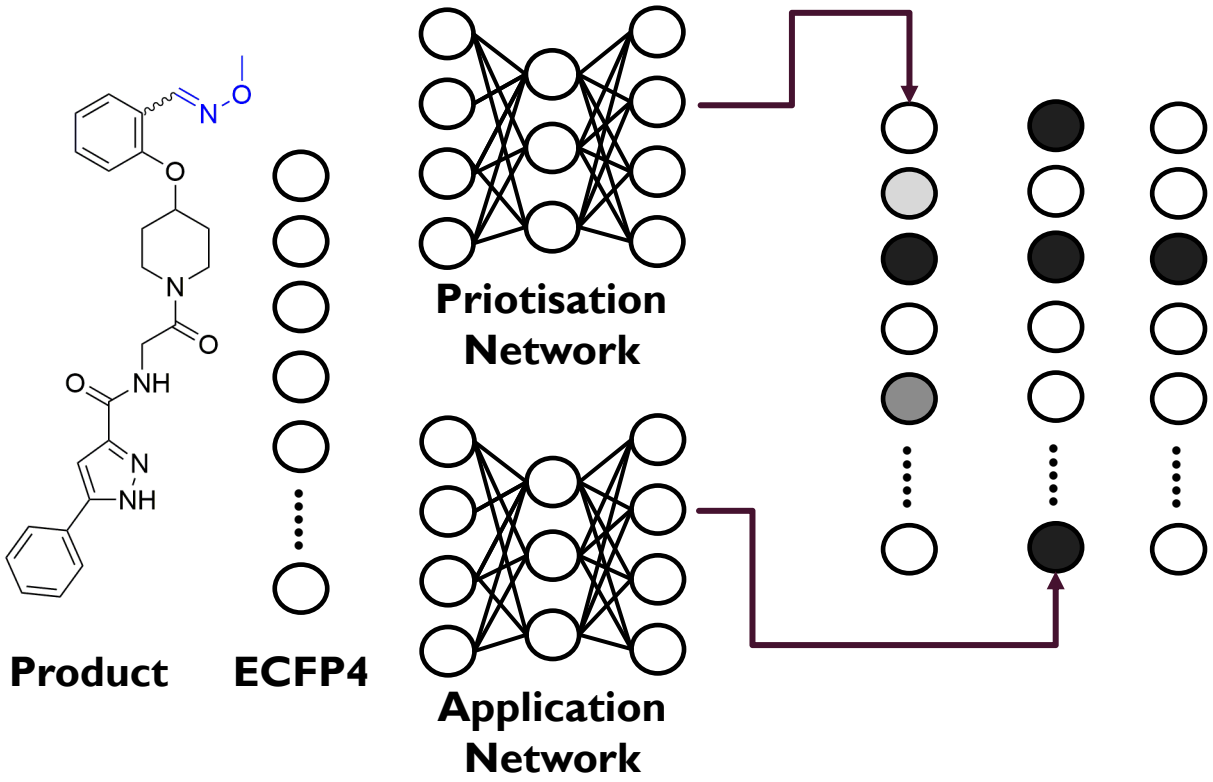




Thakkar et al., *Chemical Science*, **2020,** *11* (1), 154-168.
Bjerrum E, **Thakkar A**, MLST, **2020**, 2

IMPROVING TEMPLATE PRIORITISATION

Paper



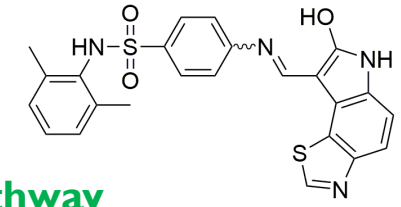
Thakkar et al., *Chemical Science*, **2020,** *11* (1), 154-168.
Bjerrum E, **Thakkar A**, MLST, **2020**, 2

PREDICTING MULTISTEP PATHWAYS

Code

Paper

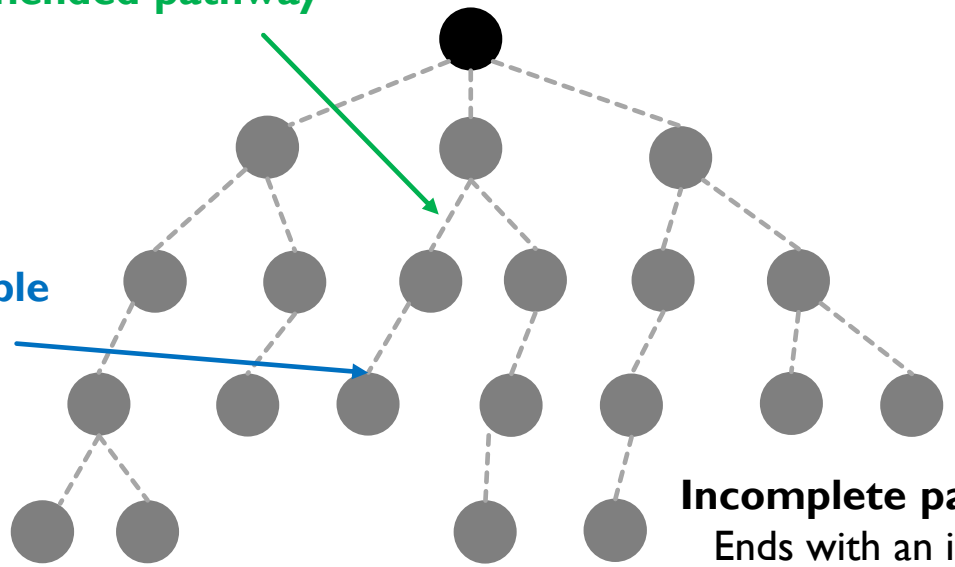
Target Compound



Recommended pathway

Commercially available building blocks

Compound cannot be synthesised



Incomplete pathway
Ends with an invalid compound

Node expansion according to:

$$UCB = \frac{Q}{N} + C \times \sqrt{2 \times \frac{\ln N_{-1}}{N}}$$

Q: sum of previous rewards
N: number of child node visitations
N_{-1}: Number of parent node visitations
C: Exploration/Exploitation hyperparameter

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

ACCESSIBILITY OPTIONS FOR END USERS

Code

Paper

GitHub Repo

- For developers
- For end users familiar with scripting and command line

README.md

Unwatch 23 | Unstar 170 | Fork 39

AiZynthFinder

license MIT | tests passing | codecov 83% | code style black | release v3.0.0 | Open in Colab

AiZynthFinder is a tool for retrosynthetic planning. The algorithm is based on a Monte Carlo tree search that recursively breaks down a molecule to purchasable precursors. The tree search is guided by a policy that suggests possible precursors by utilizing a neural network trained on a library of known reaction templates.

Prerequisites

Before you begin, ensure you have met the following requirements:

- Linux, Windows or macOS platforms are supported - as long as the dependencies are supported on these platforms.
- You have installed anaconda or miniconda with python 3.6 - 3.8

The tool has been developed on a Linux platform, but the software has been tested on Windows 10 and macOS Catalina.

Installation

For end-users

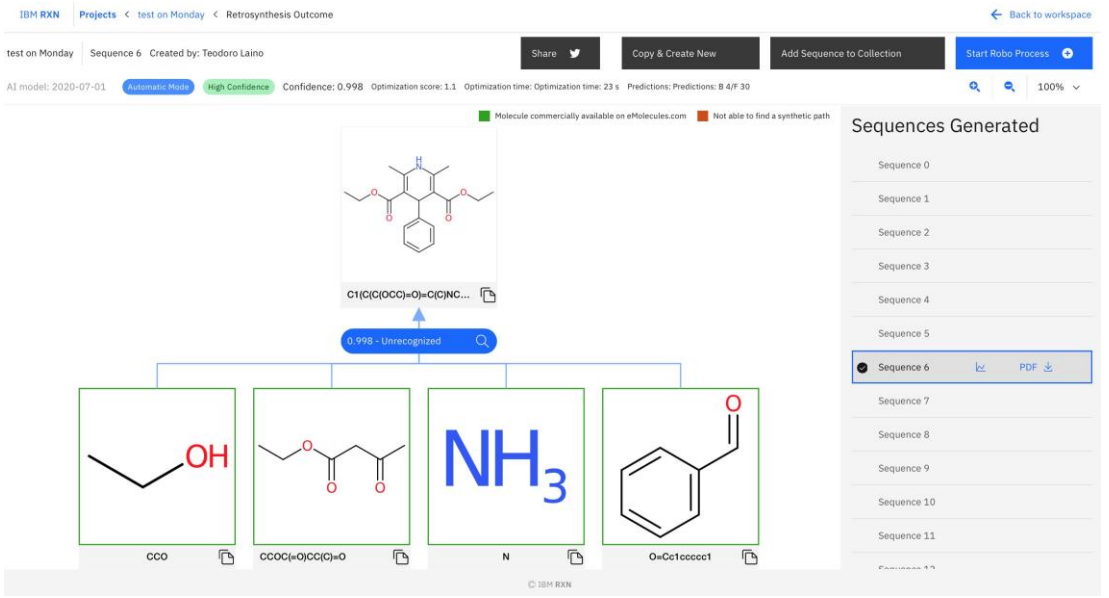
First time, execute the following command in a console or an Anaconda prompt

```
conda env create -f https://raw.githubusercontent.com/MolecularAI/aizynthfinder/master/env-users.yml
```

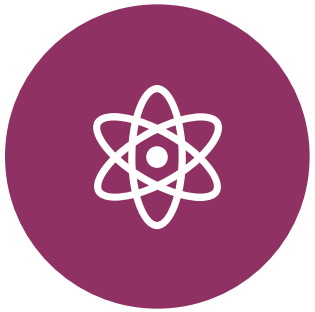
Graphical User Interface

- Deployed internally in AZ
- Allows interactive exploration
- Functionality not in repo

IBM RXN Graphical user interface below:



ACCESSIBILITY OPTIONS FOR END USERS



GDB CHEMICAL SPACE

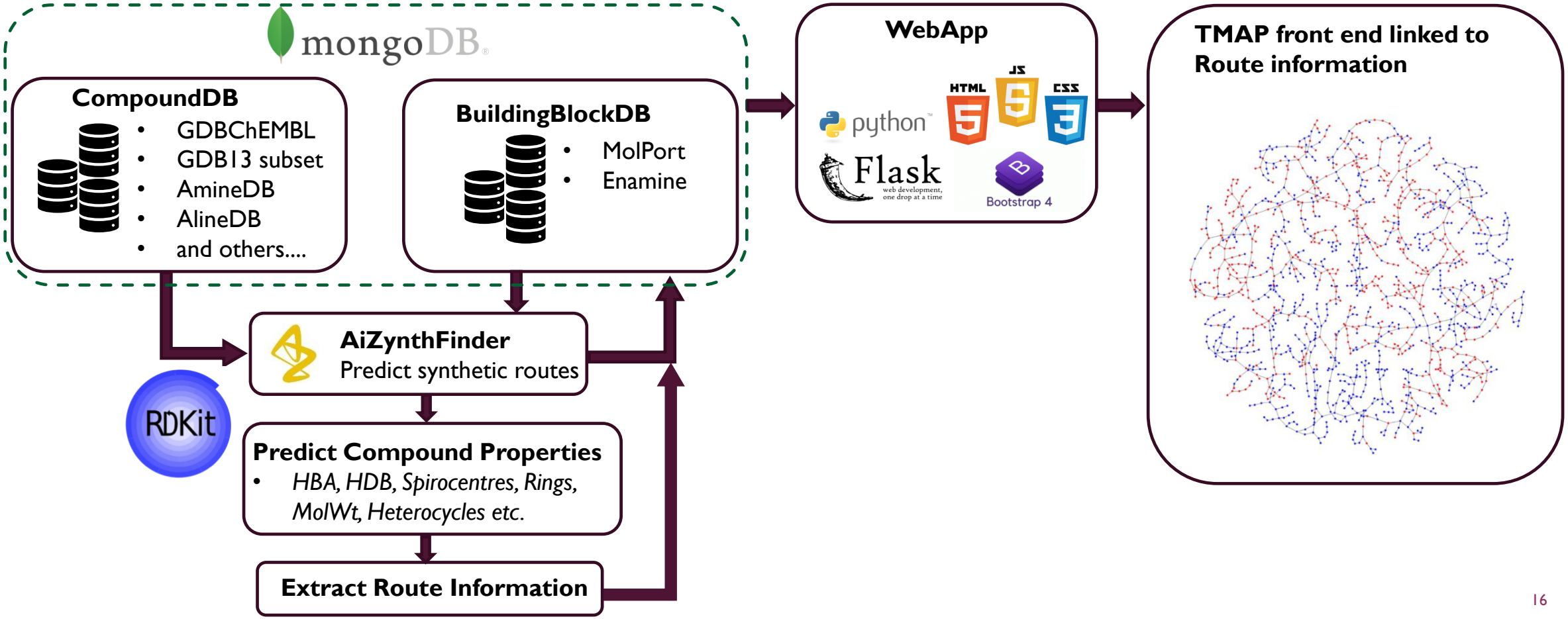
AI PLANNED SYNTHESIS

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

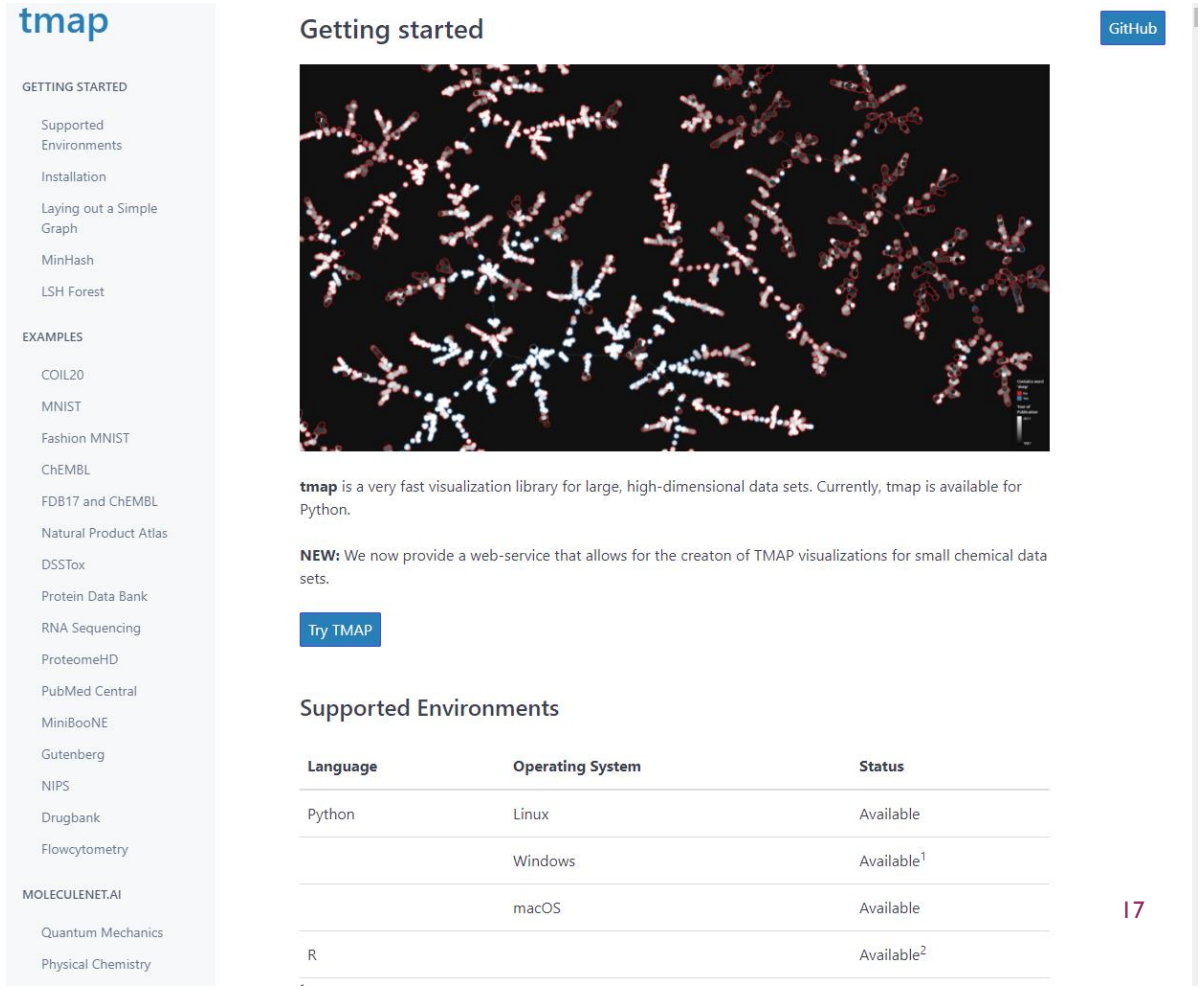
GDB ROUTE BROWSER



TMAP – VISUALISATION OF CHEMICAL LIBRARIES

- Feature vector generated using minhashing
- Index into LSH forest
- Generation of k-nearest neighbour graph
- Minimum spanning tree algorithm for layout
- Embed into 2d using co-ordinates

https://tmap.gdb.tools/



D. Probst, J.-L. Reymond, Visualization of very large high-dimensional data sets as minimum spanning trees. *J Cheminform* **12,** 12 (2020).

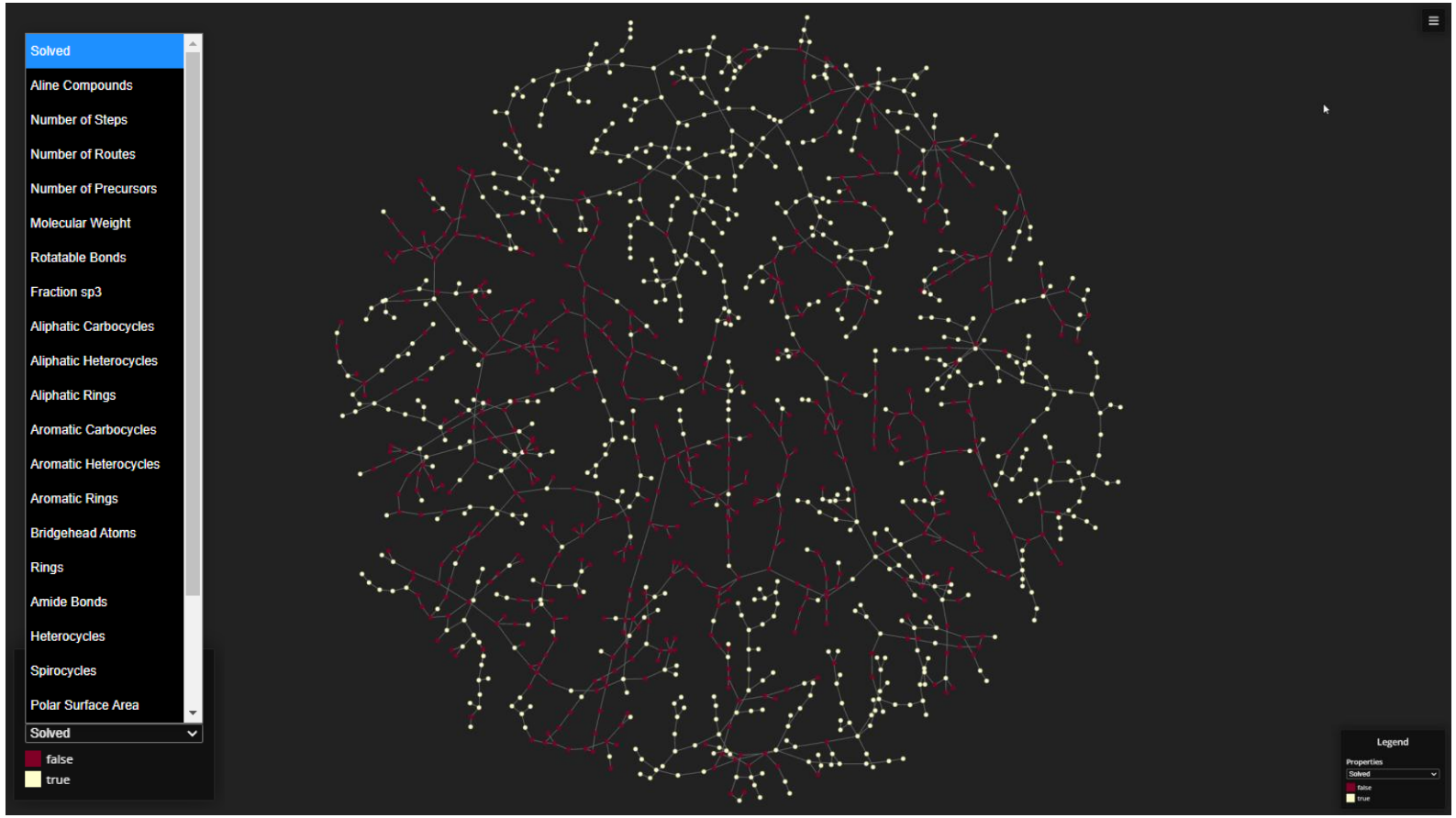
GDB RETROSYNTHESIS

Dataset	Compounds	% Solved by AiZynthFinder	Number of Routes	Number of Steps	TMAP Created
AmineDB	44,929	17.7	427,493	2,829,675	Y
DiamineDB	1,323	56.2	12,207	59,987	Y
GDB13_ABCDEFGH	994,840	19.7	8,501,323	54,933,389	N
GDBChEMBL_A	1,490,508	33.8	12,826,692	75,029,188	N
Total	2,531,600	28.0	21,767,715	132,852,239	

Reaxys®

>54M

chemical reactions



COLORATION BY PROPERTIES

Diamine Database

Number of rings: Max 2

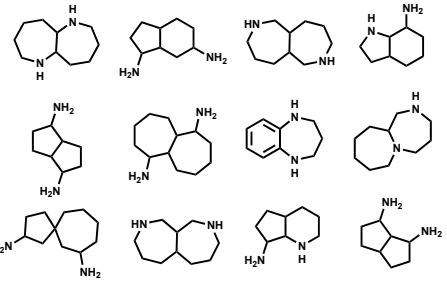
Ring Sizes: 5, 6, and 7

Number of Amines: Max 2

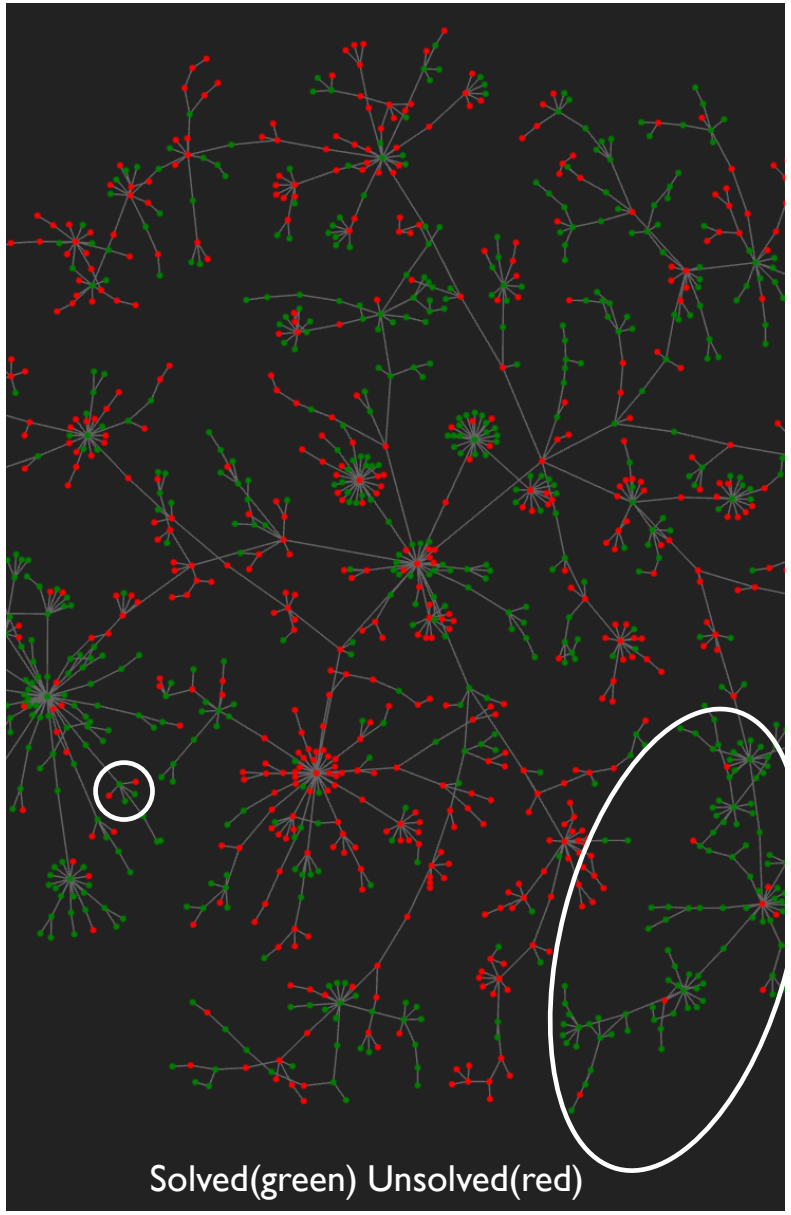
Exo- and Endocyclic

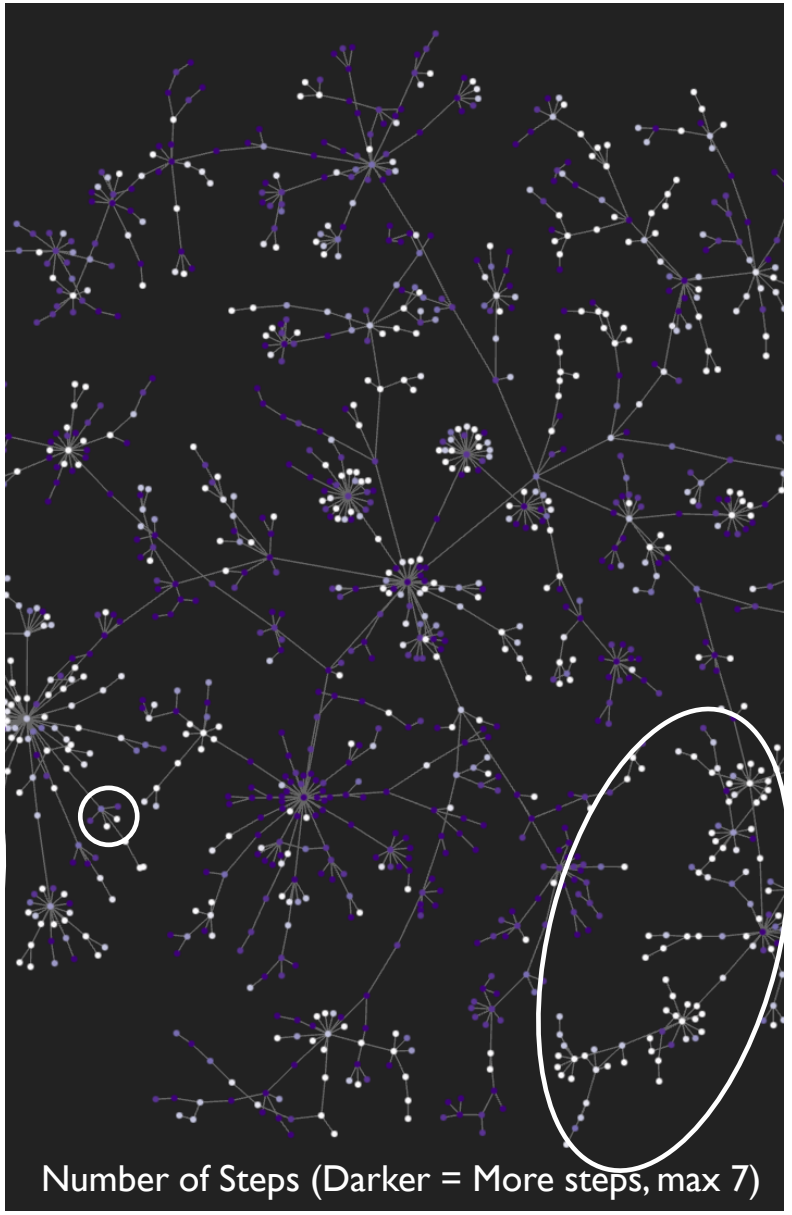
1323 Molecules

57 % Novel



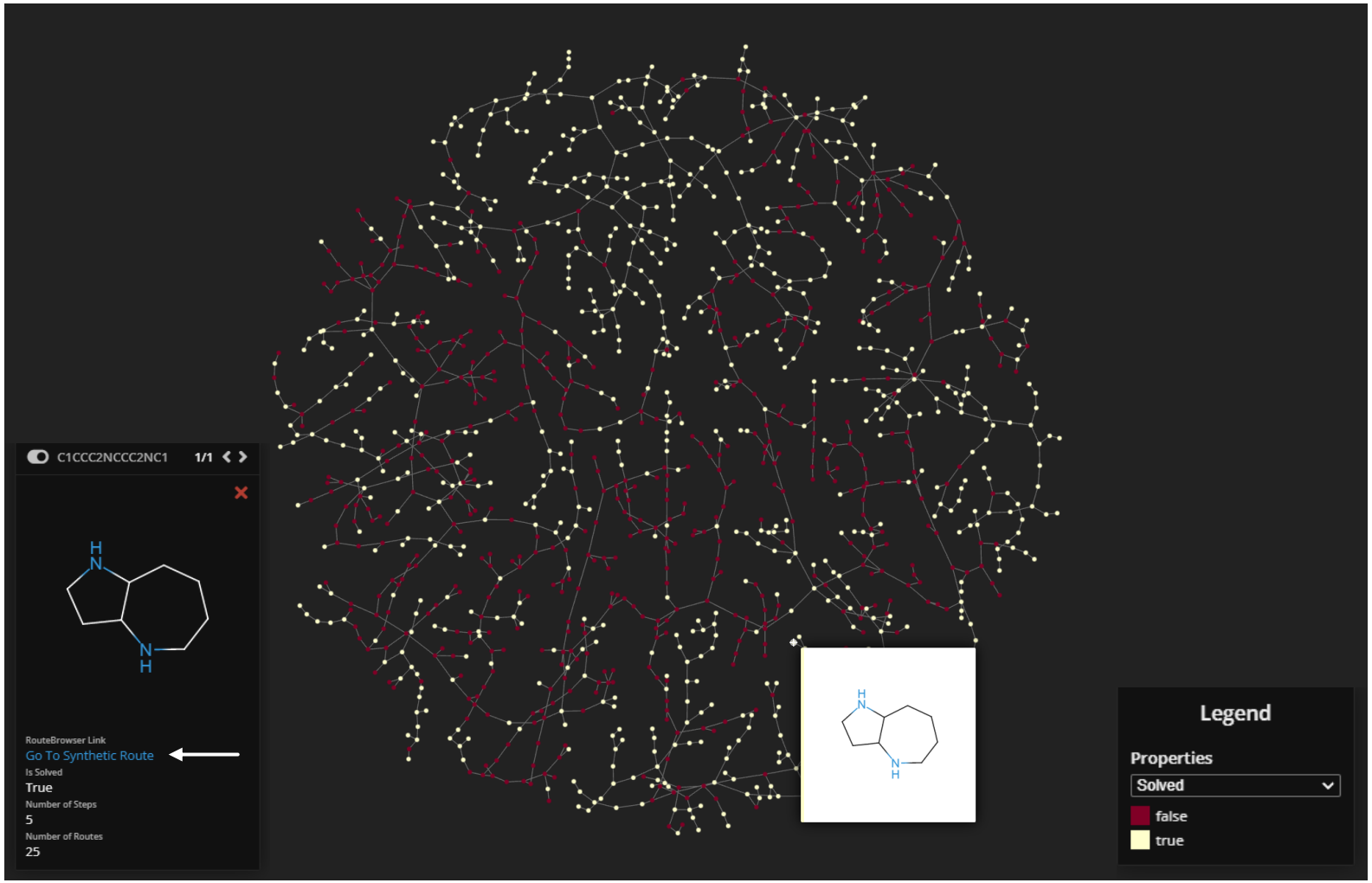
- Helps prioritisation of compounds
- View of related compounds
- Relationships between compounds in the library





ENABLING A RICHER UNDERSTANDING OF CHEMICAL SPACE

- Short synthetic routes to privelaged scaffolds
- Chemists know the reactions for disconnection
- AiZynthFinder knows the reactions and the building blocks available



PRECOMPUTED SYNTHETIC ROUTES FOR EACH COMPOUND

GDB Route Browser

Home Login Register

SMILES, InChI or InChI Key

Compound Identifiers:

SMILES: C1CCC2NCCC2NC1

InChI: InChI=1S/C8H16N2/c1-2-5-9-8-4-6-10-7(8)3-1/h7-10H,1-6H2

InChI key: SLCIXBLNPIPMOC-UHFFFAOYSA-N

Route Details:

25 routes predicted in 180.65 seconds

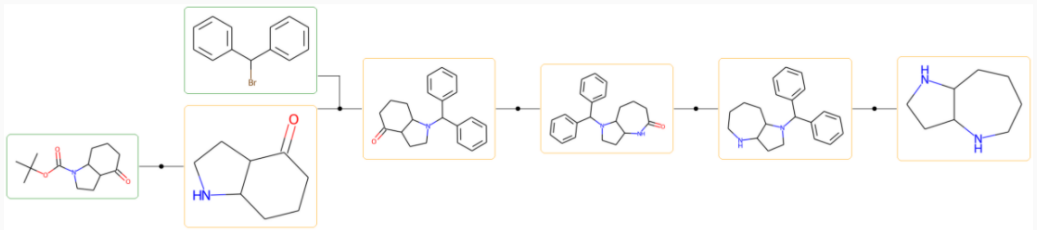
Currently showing: Route 1 of 25

Number of Steps: 6

Number of Precursors: 3

Predicted Route:

***Compounds highlighted in a green box are commercially available**



Like Dislike

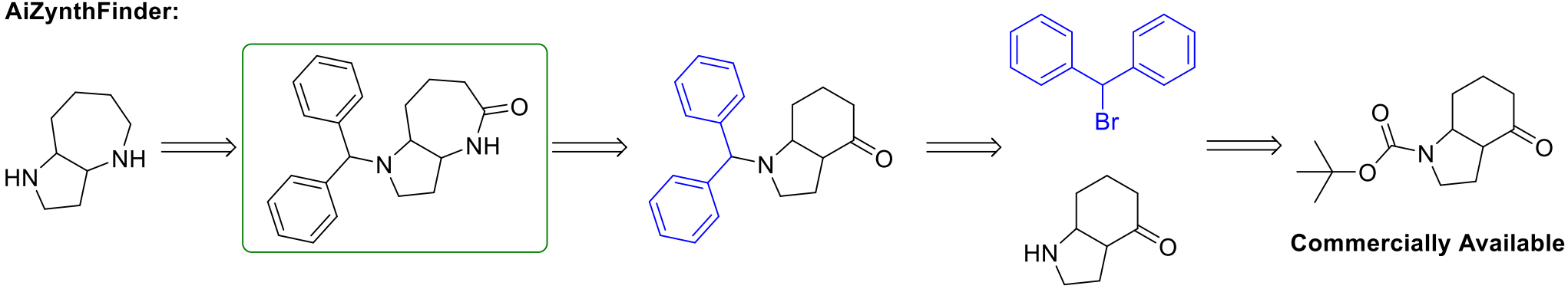
Precursors Available:

APPLICATION TO GDB – EXPERIMENTAL VALIDATION

ASKCOS: No Route Found

IBM RXN: No Route Found

AiZynthFinder:



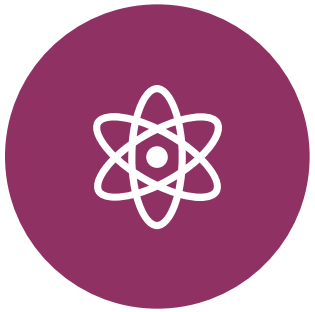
KEY STEP:
Beckmann Rearrangement
-Precedented on similar motif

Interchangable Protecting Group

Commercially Available

Successfully synthesized in the lab!

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, calcultion of properties, synthesis prediction, prioritisation

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

SUMMARY

- AiZynthFinder – An Open Access Retrosynthetic Planning Tool
- Established methods for improving AiZynthFiner using data augmentation and specialised models.
- Built specialised tools to target specific needs arising from chemists.
- The tools have demonstrated project impact as shown for the case of GDB.
- Nearly everything is open source and published open access for the community!

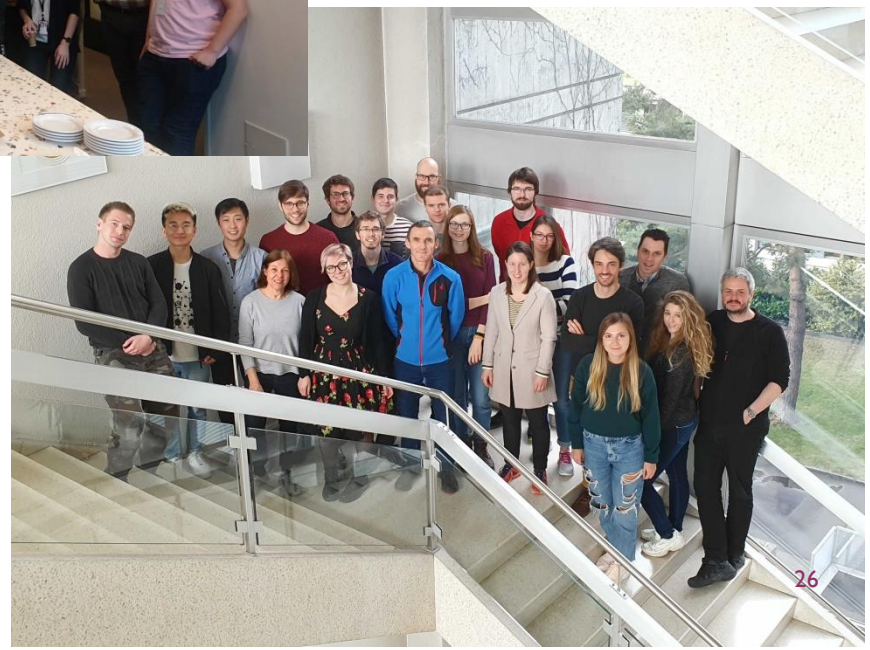
- Linking visualisation to chemical synthesis planning will be released in the near future (currently writing and refactoring the code)

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.



Feedback or Questions?

Email:
amol.thakkar@unibe.ch