

Data foundations

Developing AI agents for
laboratory science in pharma

Nessa Carson

*Predictive Science Digital & Automation, Pharmaceutical Sciences,
R&D, AstraZeneca, Macclesfield, UK*

15 April 2026



Digital transformation

Digitalization is one of the most prominent issues for chemical industry (according to CEO's)

- 👕 For *growth...*
- 👕 For *making the future happen...*
- 👕 For *sustainability...*

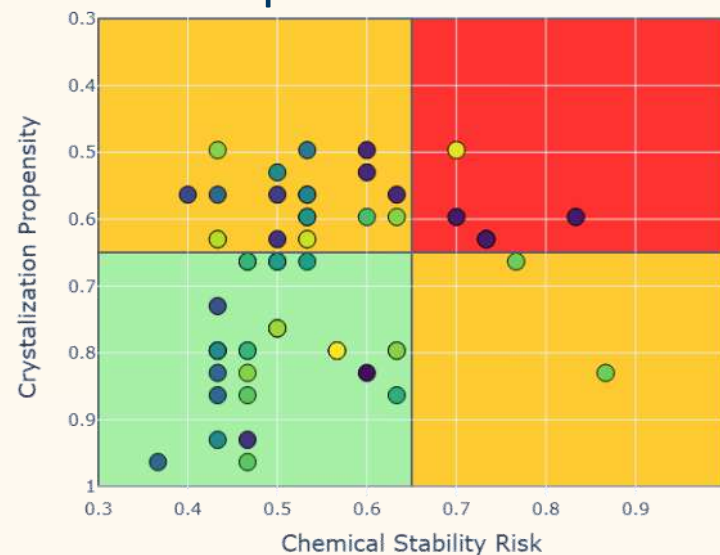


Transformative change

Smart sensors in the lab



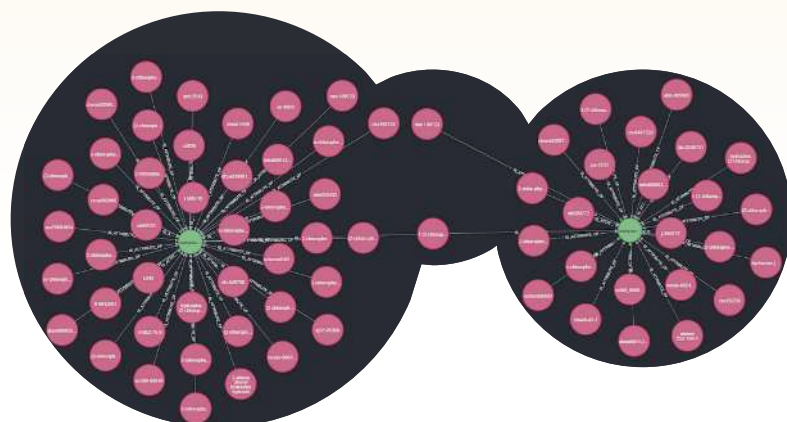
Software to enable experimentation



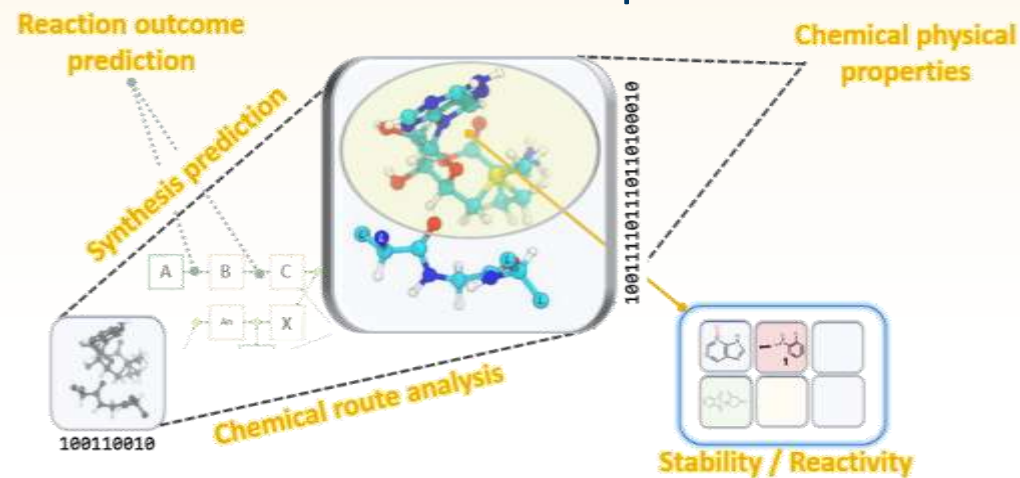
Autonomous robotics



Big chemistry data



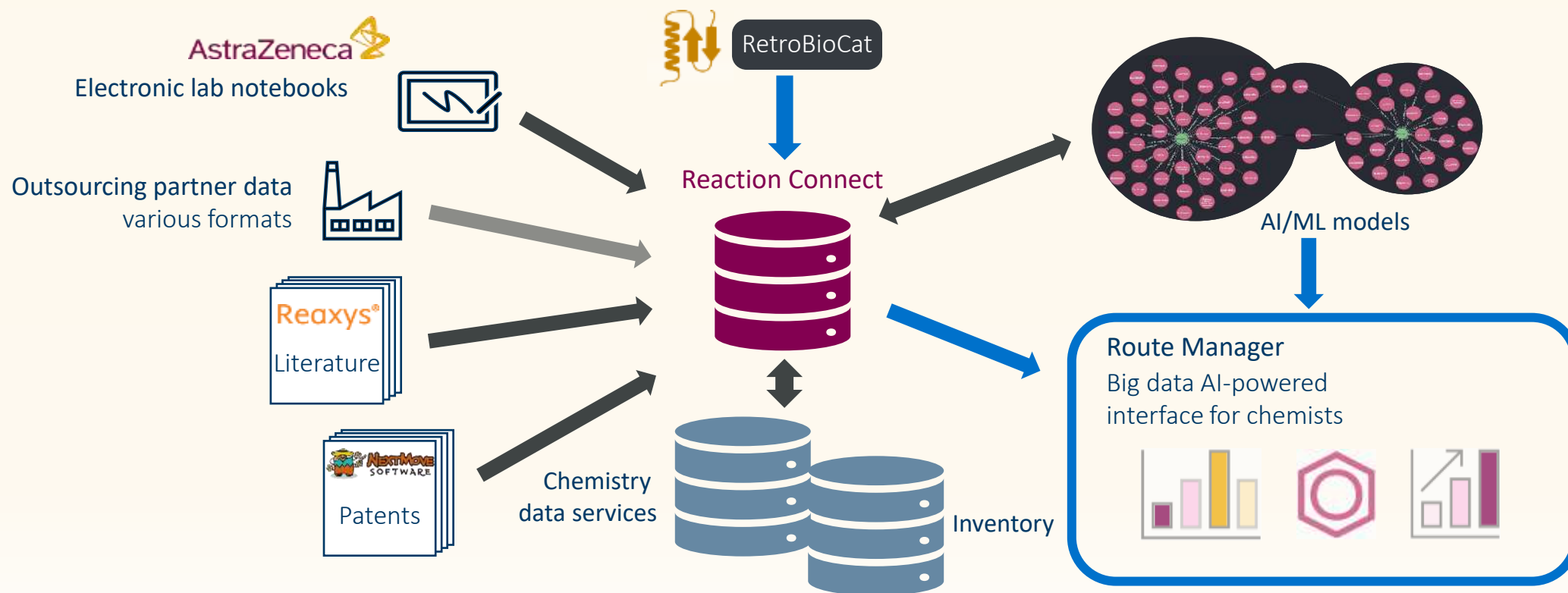
Accessible computation



The bright future
needs our data from traditional labs



Data flow enables scientific innovation



Data chemists want:

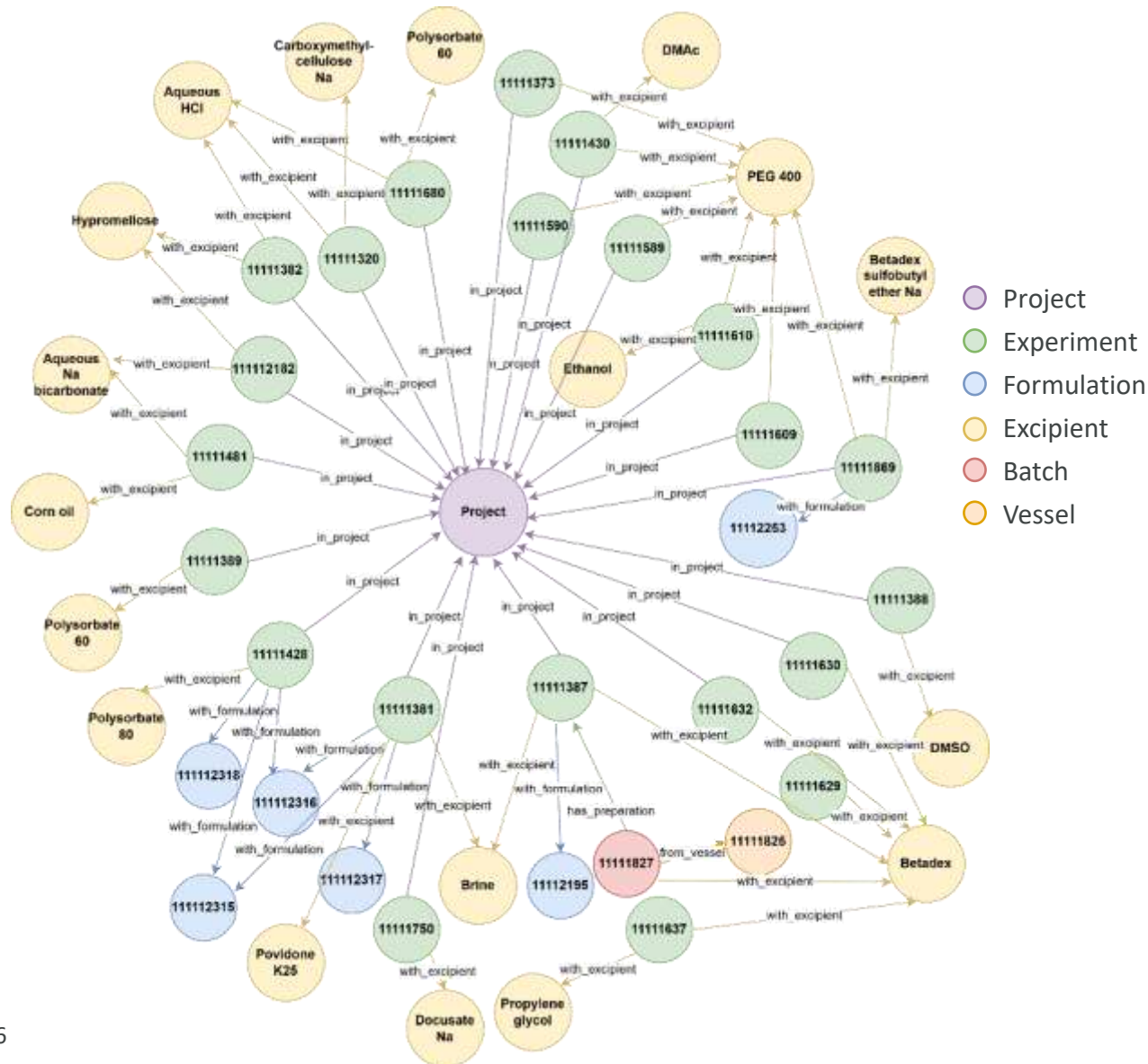
- Good data standards
- Incorporation of FAIR metadata

Lab chemists want:

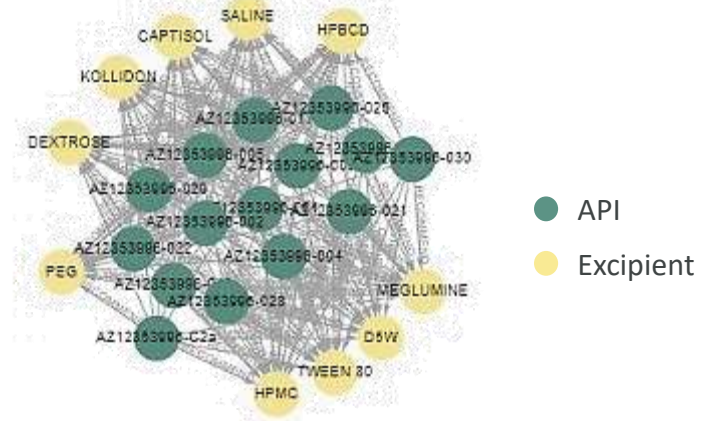
- Decrease in admin
- Incorporation of FAIR metadata



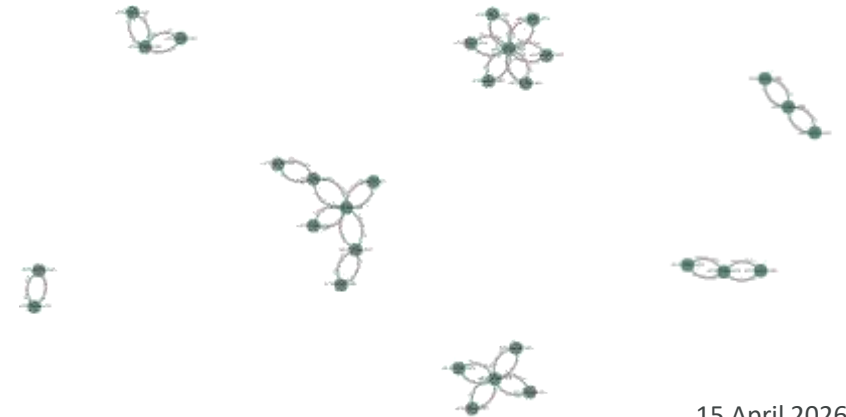
We do amazing things with big data



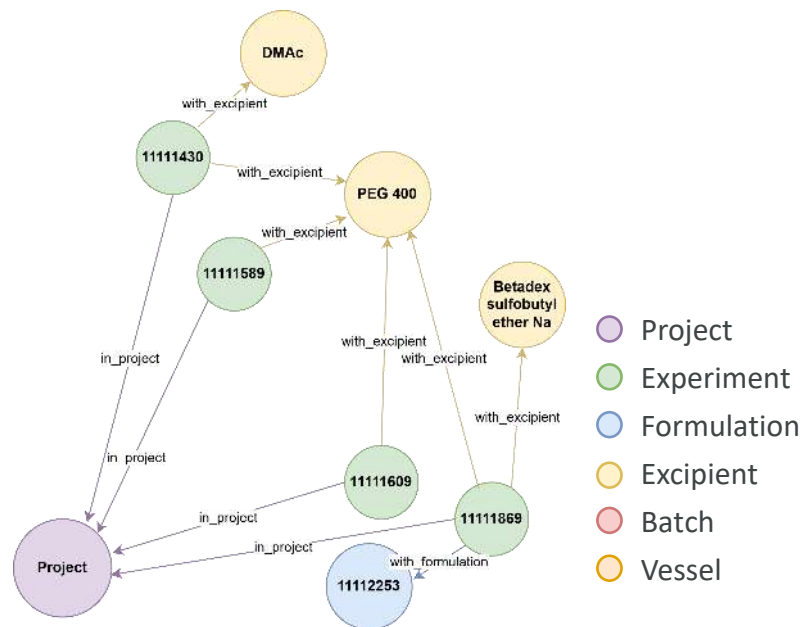
Explore excipient recommendations



APIs with similar physicochemical properties



The landscape without CxO data



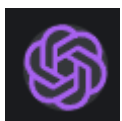
- AZ Early Chemical Development outsources **70–80% of their *total chemistry*** to CRO's/CDMO's
- Outsourced documentation comes in with... **mixed quality**
- Main challenges:
 1. Unstructured reports
 2. Formats difficult for computers to read
 3. Embedded images of structures & reactions
 4. Knowledge transfer to *a few* people



Can't we just "AI it all"?

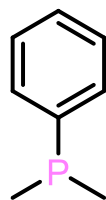
Easy mode

"Propylene carbonate (1.0 L, 1.204 kg) was then charged to the reactor"

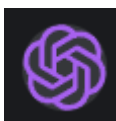


"The volume of propylene carbonate used is 1.0 litre"

Hard mode



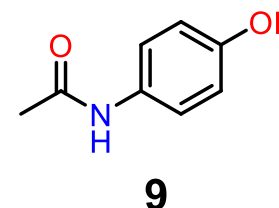
"2.5 mol% of the phosphine ligand was used"



"It seems that 2.5 mol% of a phosphine ligand was used. It is likely to be Me₂PPh"

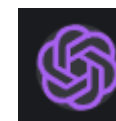
Nightmare mode

Page 1:



Page 6:

"Compound 9 was formed in 89% yield"

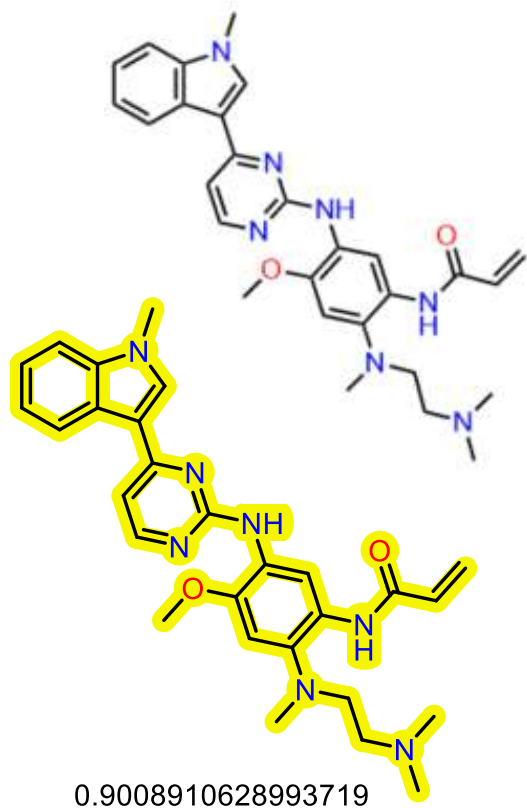


"I cannot tell what Compound 9 is — it's just a label used in the report"

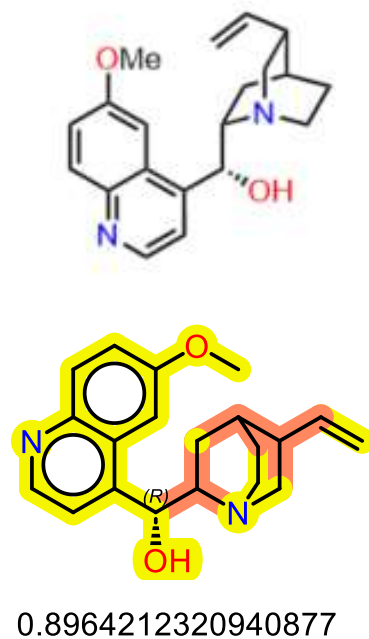


Optical structure recognition

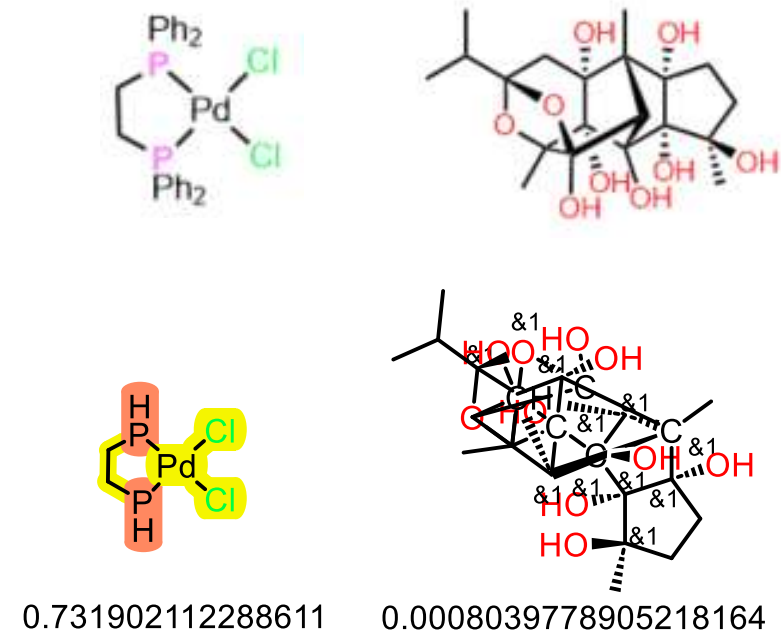
Easy mode



Hard mode



Nightmare mode



Bespoke AI tools for our chemistry needs

- Chemical data are **always challenging**

Computers
understanding
structures

Lack of
standardization

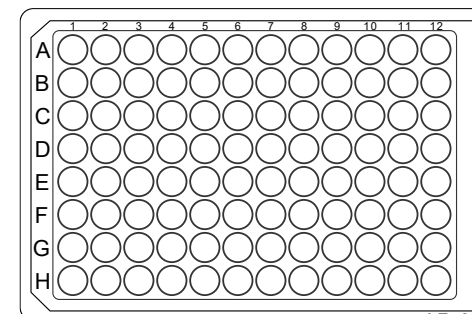
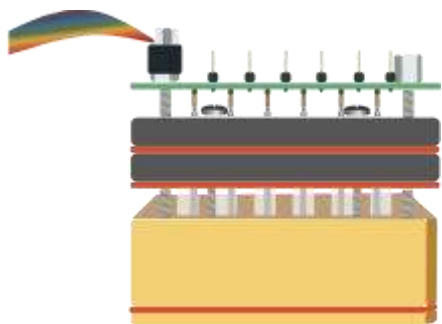
ML methods for the
small data regime

Contents (& quality!) of
each dataset/reaction
vary wildly

Goals

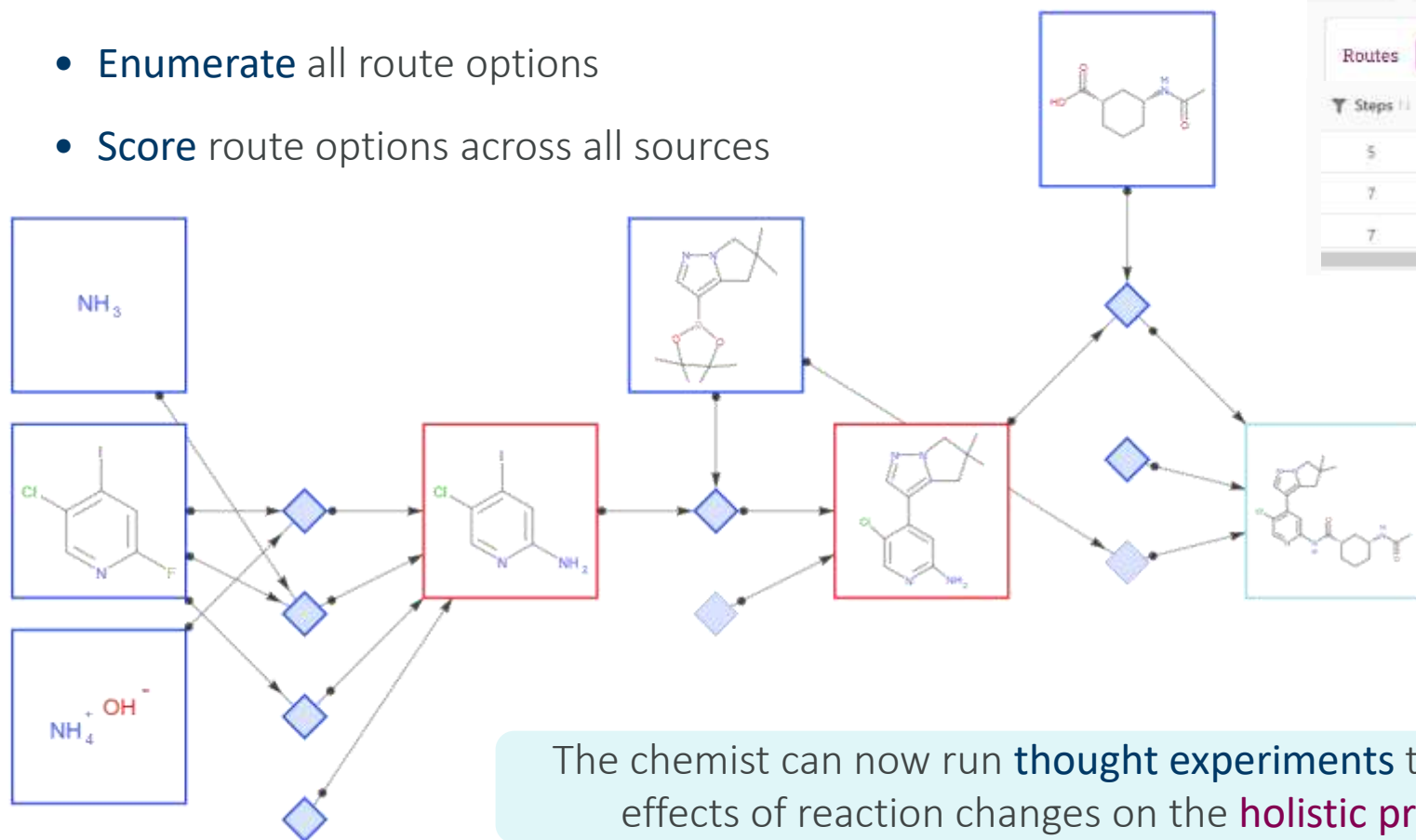
- Extract **information** from documentation, for reuse in **ML** and the **Route Manager** interface
- **Impeccable** accuracy – but accept high error % early on

We started with internal historical data from **high-throughput experimentation**



We do amazing things with big data: Route Manager

- Retrieve and display reaction data from our **ELN's** and **external data** sources
- **Enumerate** all route options
- **Score** route options across all sources



The chemist can now run **thought experiments** to see the effects of reaction changes on the **holistic process**

Route metrics

Details Routes

Routes **Generate** Total: 9 of 1000 Annotated: 0 Selected: 0 Pareto **Show/Hide Columns**

Steps	LLS	CDG	pCOG	PMI	aPMI	pPMI	Overall Yield
5	5	532	5.69	152	152	171	41
7	7	144	9.65	167	167	284	24.9
7	7	239	9.65	159	159	284	29

Editing the data

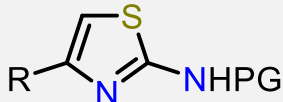
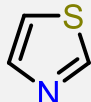
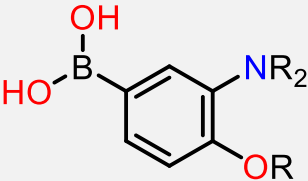
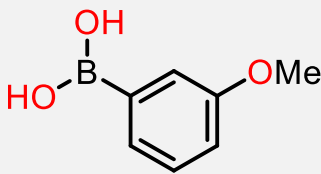
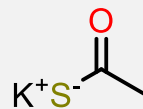
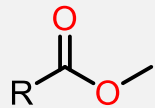
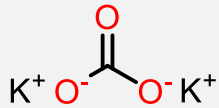
Products

Product id	AVIWDYS/SPOOAR-LSDHHAJUNA-N
Compound name	
Product name	(1S,3R)-3-acetamido-N-(5-chloro-4-(5,5-dimethyl-1H-imidazol-2-yl)pyridin-2-yl)propanamide
Sample id	
Reagent id	N/A
Yield (%)	26
Mw (g/mol)	430
Total mass (g)	0.0568



Working with the agent

- It makes a lot of mistakes, at first!

What the author meant	What the agent saw	What the agent “understood”
	“thiazole”	
	“Ar-B(OH) ₂ ”	
	“KSCOMe”	“???”
	“ester”	



Working with AI tools

- Any AI tools work best with **more relevant information**

Curated
chemical lists

Controlled, high-accuracy lists of
catalysts, solvents, reagents

Algorithmic
tools

Understanding structures based on the
chemical name (including common misspellings...)

Synonyms

From inventory, materials management, vendors,
med chem databases, reliable external sources...

And as much existing information on the chemistry
as we can get!

Learnings from multiple AI projects

- It's pointless to hold the AI's hand... but **in-depth scientific guidance** is critical
- Everything in chemistry is **complex** from a data perspective
- Accept that this **takes work** and will be far – perhaps very far – from perfect straight away
- If we can get more knowledge out of a tool, **futureproof its results** by doing that!
- Be flexible to **expand the existing data structure**
- Always use **FAIR** data processes
- Work with outsourcing partners to improve **how we source data** in the first place.
But that isn't so simple!



Acknowledgements

Predictive Science DS

Christof Jäger
Per-Ola Norrby
David Buttar
Megha Anand
Lucy Arvidsson
Christoph Bauer
Felix Faber
Arseny Kovyrshin

Cata-List team

Per-Ola Norrby
Anders Egnéus
Will Lander

AI agent team

Christof Jäger
Felix Faber
Megha Anand
Mikhail Kabeshov
Daniel Maddox
Maliha Sultana

AI in Chemistry conference:

<https://www.rscbmcs.org/events/aichem9/>

Selected scientific & digital resources:

<https://supersciencegrl.co.uk/links/>



Confidentiality Notice

This file is private and may contain confidential and proprietary information. If you have received this file in error, please notify us and remove it from your system and note that you must not copy, distribute or take any action in reliance on it. Any unauthorized use or disclosure of the contents of this file is not permitted and may be unlawful. AstraZeneca PLC, 1 Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge, CB2 0AA, UK, T: +44(0)203 749 5000, www.astrazeneca.com

