Creating a Reaction Knowledge Graph with Open-Source Software

Christos Kannas [1], Tomas Bastys [1], Thierry Kogej [1]

[1] MolecularAI, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden

12 - 14 Oct 2022

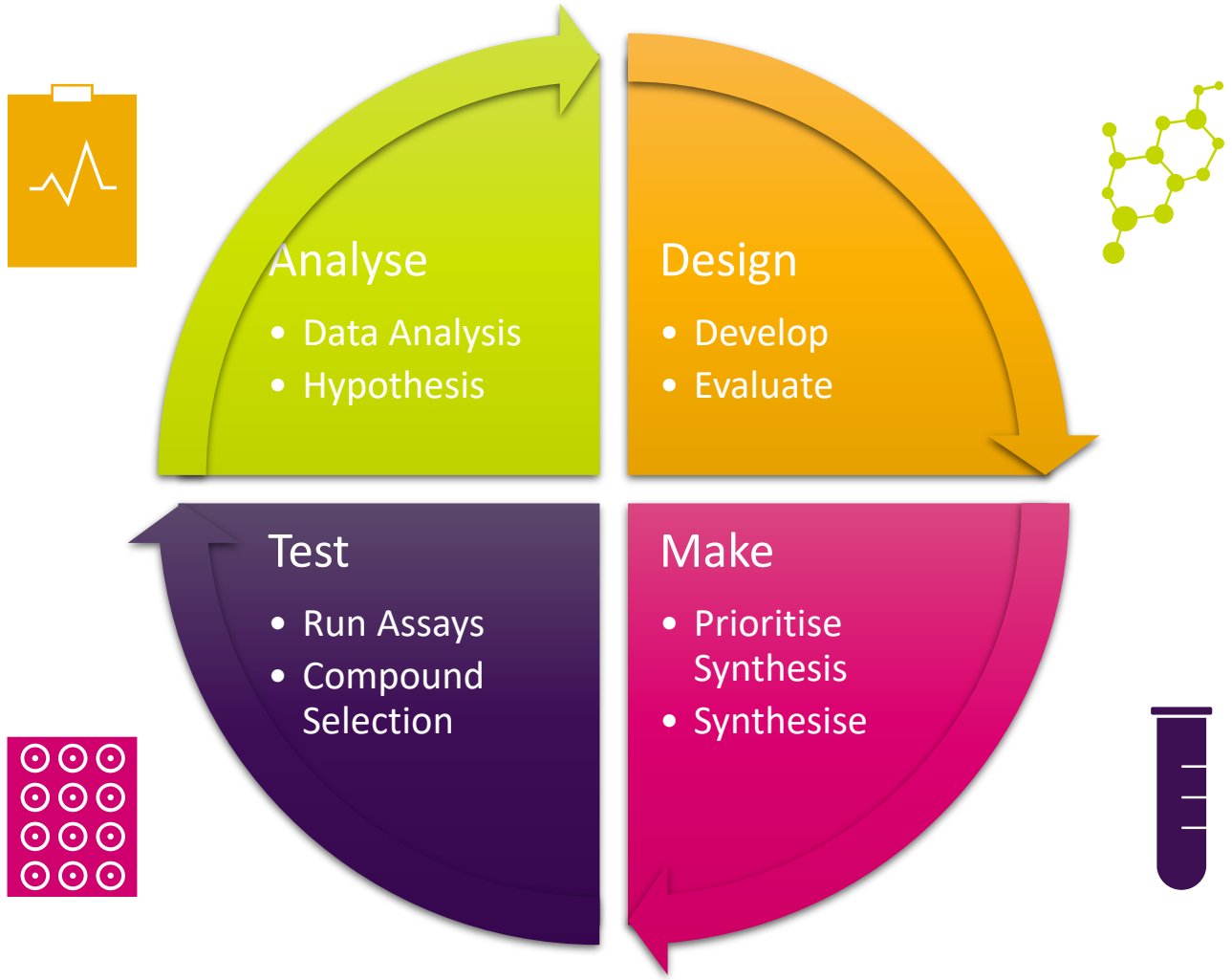
Outline

1. Background
2. Building the graph
3. Using the graph
4. Future Work

1

Background

Drug Discovery Process



Objectives

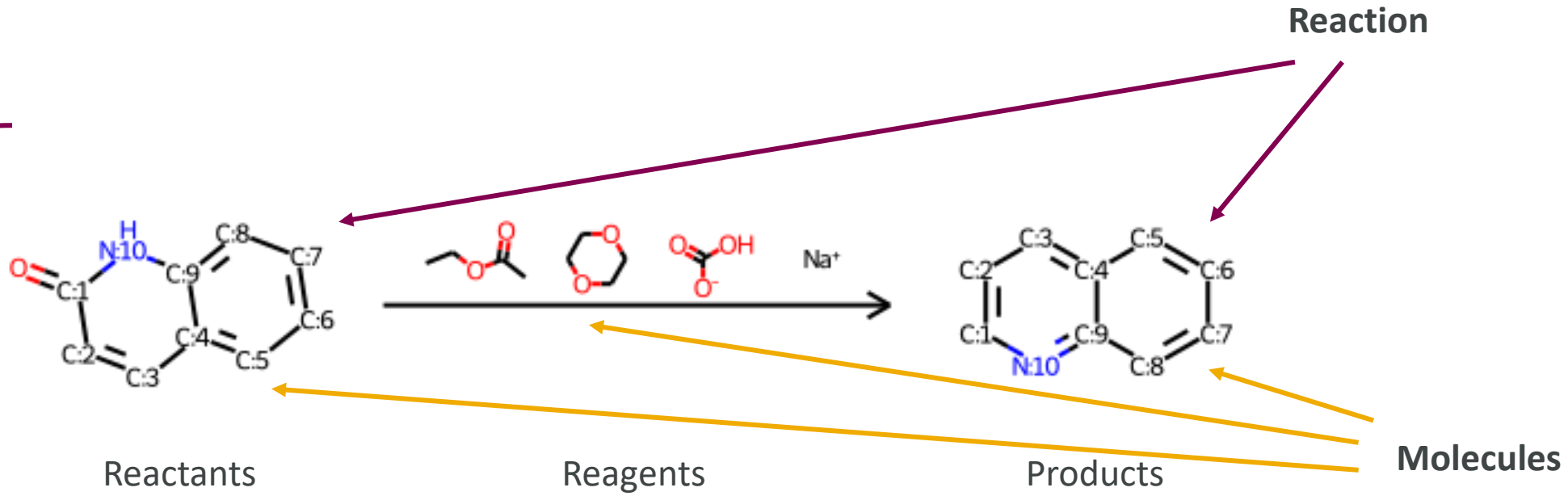
Predict Novel Reactions

Explore Reaction Space

Explore Reactants, Reagents, Products Space

Assist Synthesis Prediction

Reaction Record Example



Reaction Variation

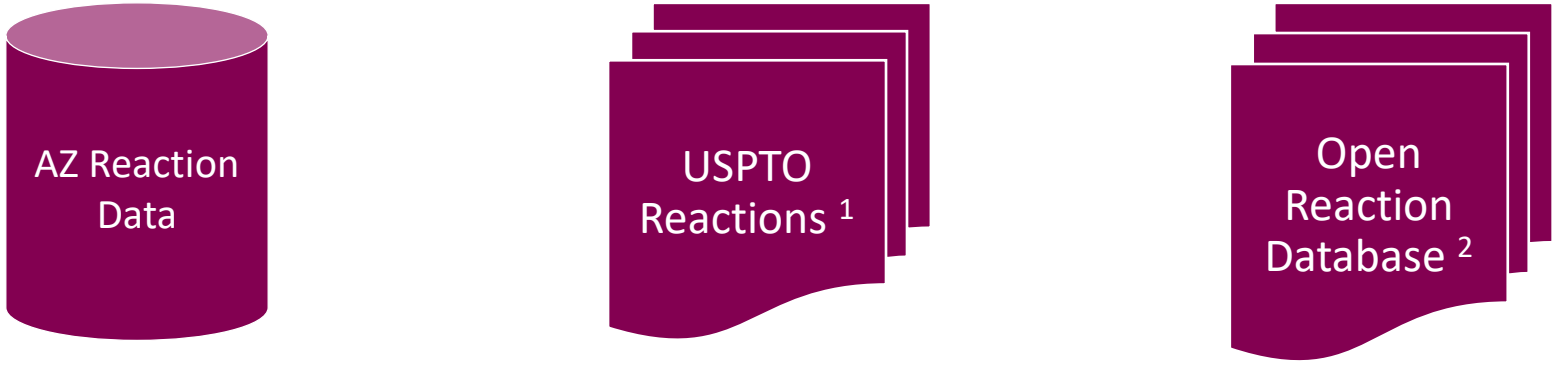
- Reaction Identifier
- Source Name
- Date
- Reaction Yield

- Classification Id
- Classification Name

Reaction Classification

USPTO Identifier: US20120270775A1;94;2012

Reaction Data Sources

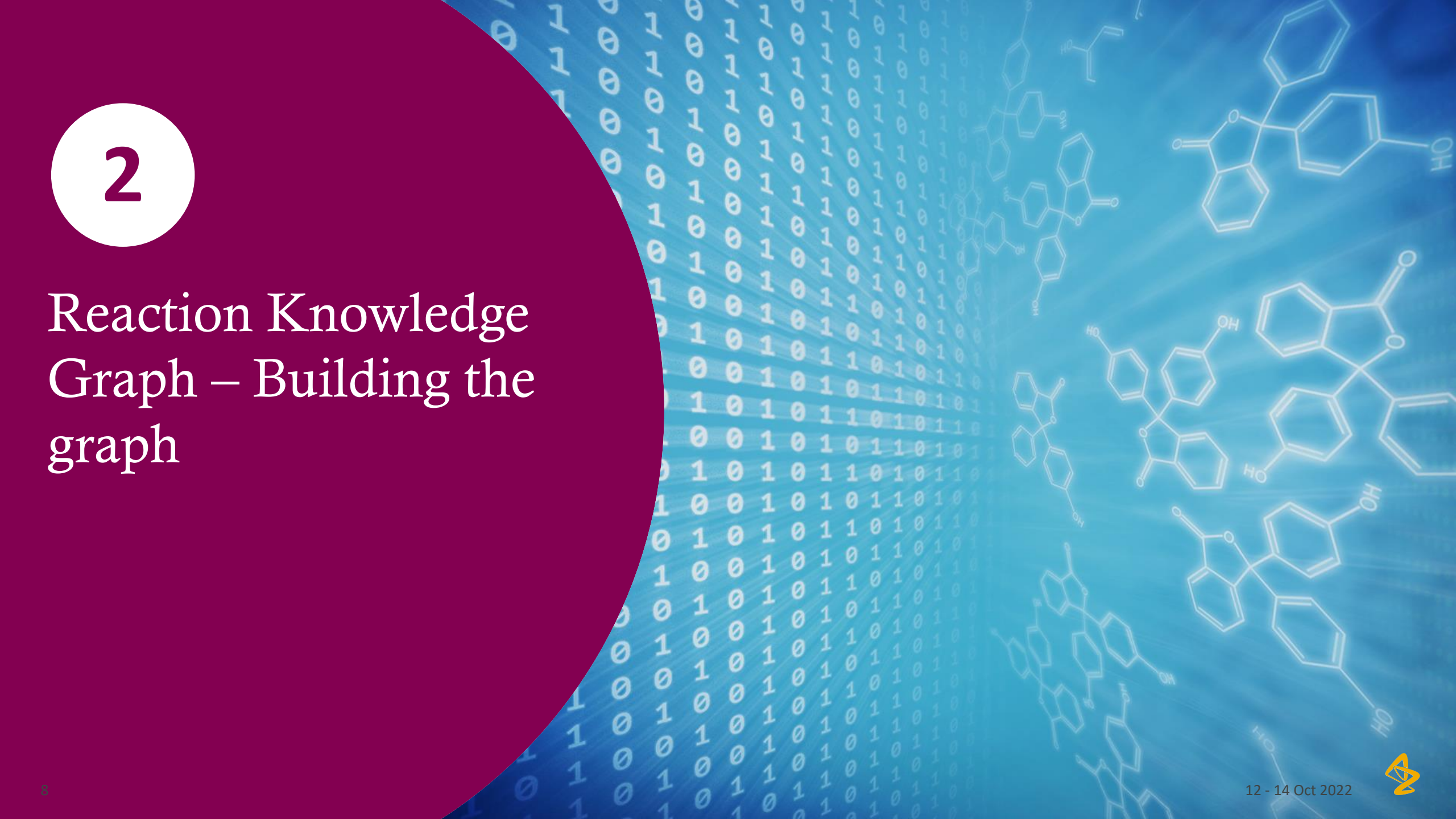


US Patents & Trademark Office

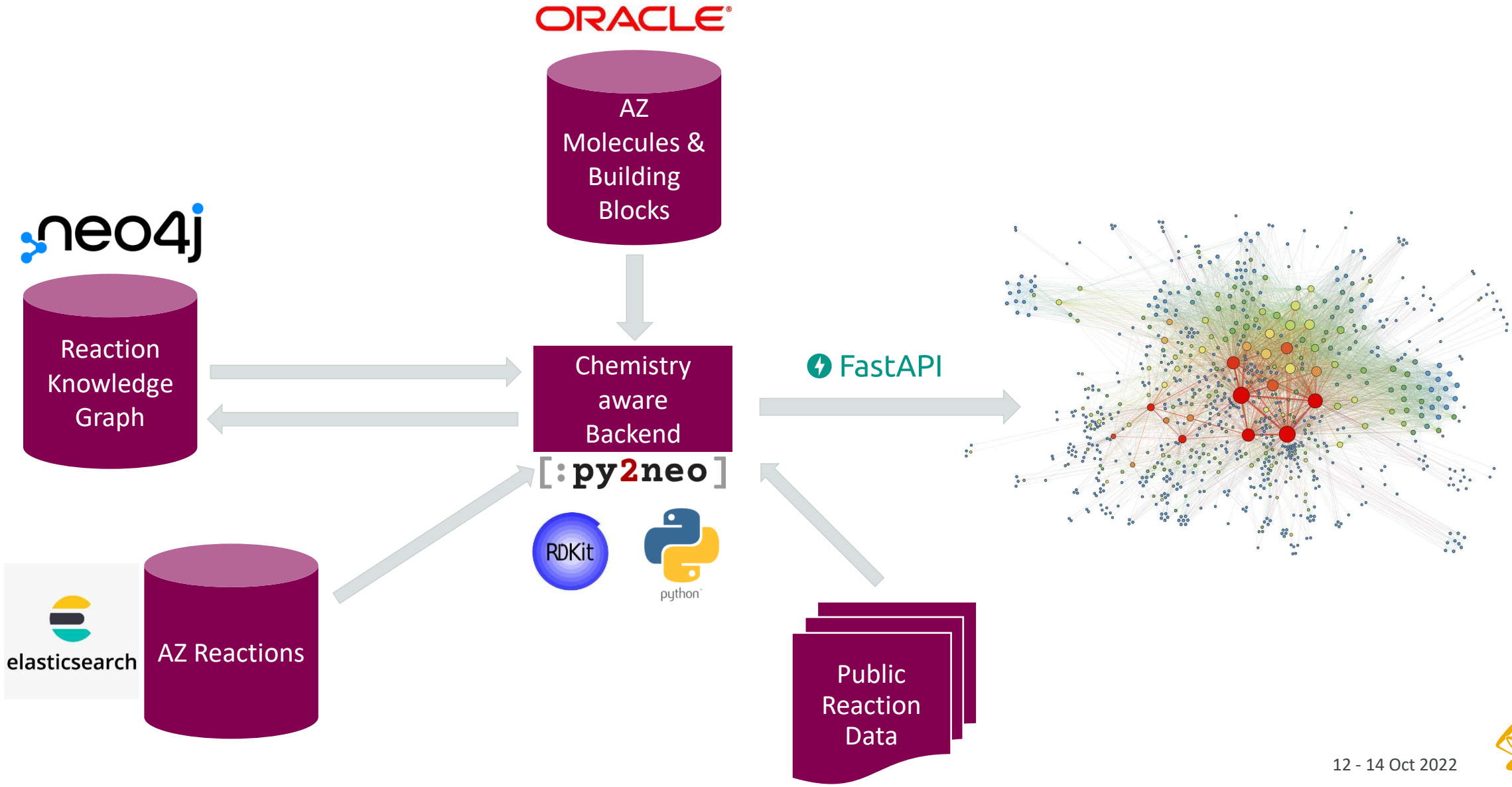
(1) Lowe, D. Chemical Reactions from US Patents (1976-Sep2016), 2017. https://doi.org/10.6084/m9.figshare.5104873.v1.
(2) Kearnes, S. M.; Maser, M. R.; Wleklinski, M.; Kast, A.; Doyle, A. G.; Dreher, S. D.; Hawkins, J. M.; Jensen, K. F.; Coley, C. W. The Open Reaction Database. J. Am. Chem. Soc. 2021, 143 (45), 18820–18826. https://doi.org/10.1021/jacs.1c09820.

2

Reaction Knowledge Graph – Building the graph



Reaction Knowledge Graph – Architecture



Reaction Data ETL pipeline

Extract

- Get data from flat files

Transform

- Validate & select reaction records
- Prepare data for loading

Load

- Generate nodes & relationships
- Store nodes & relationships to Reaction Graph DB

(1) Kannas, C.; Thakkar, A.; Bjerrum, E.; Genheden, S. Rxnutils – A Cheminformatics Python Library for Manipulating Chemical Reaction Data. 2022. https://doi.org/10.26434/chemrxiv-2022-wt440-v2.
(2) Thakkar, A.; Kogej, T.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. Datasets and Their Influence on the Development of Computer Assisted Synthesis Planning Tools in the Pharmaceutical Domain. Chem. Sci. 2019, 11 (1), 154–168. https://doi.org/10.1039/c9sc04944d.
(3) Coley, C. W.; Green, W. H.; Jensen, K. F. RDChiral: An RDKit Wrapper for Handling Stereochemistry in Retrosynthetic Template Extraction and Application. J. Chem. Inf. Model. 2019, 59 (6), 2529–2537. https://doi.org/10.1021/acs.jcim.9b00286.

Molecule Enrichment Pipelines

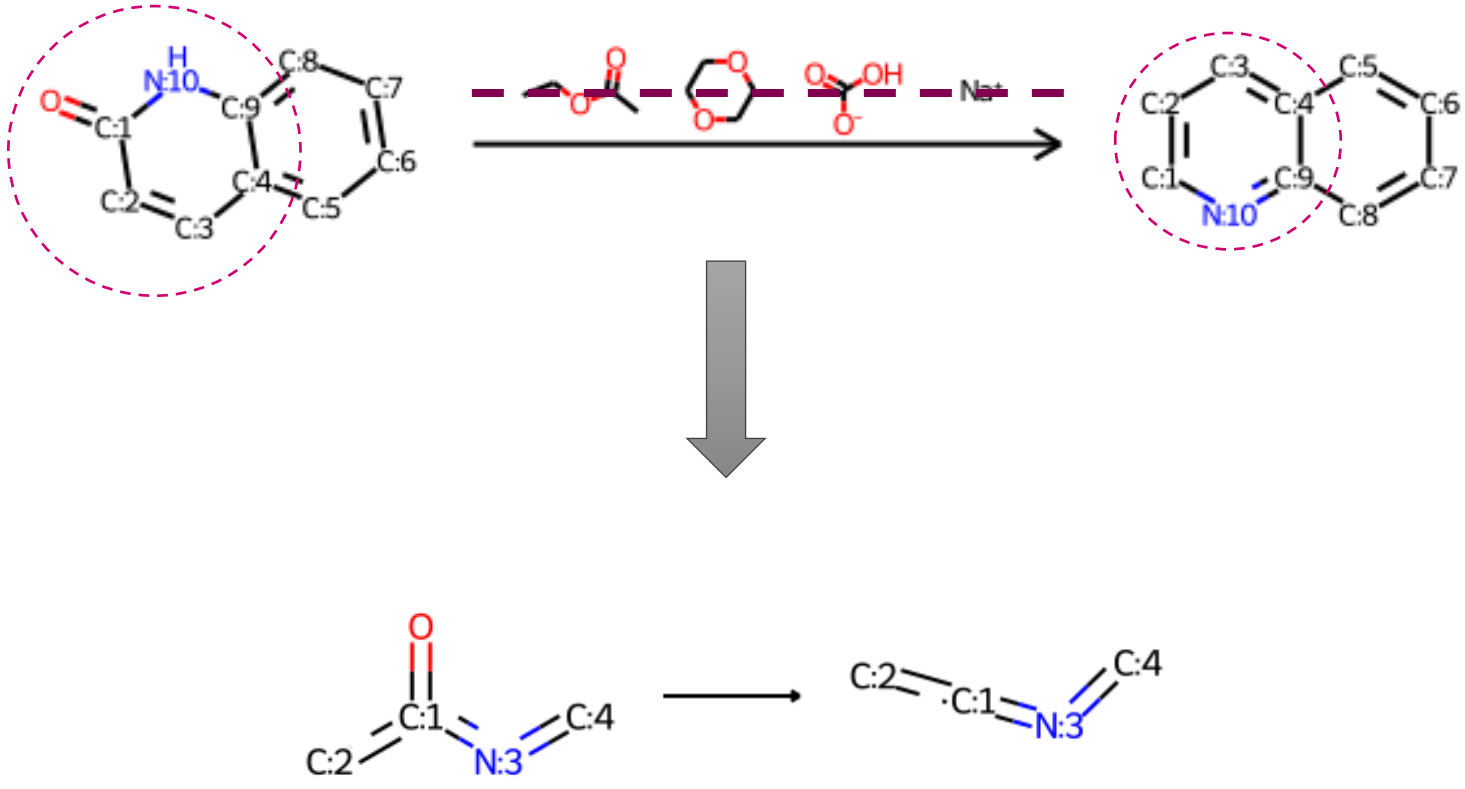
Compound Identifiers

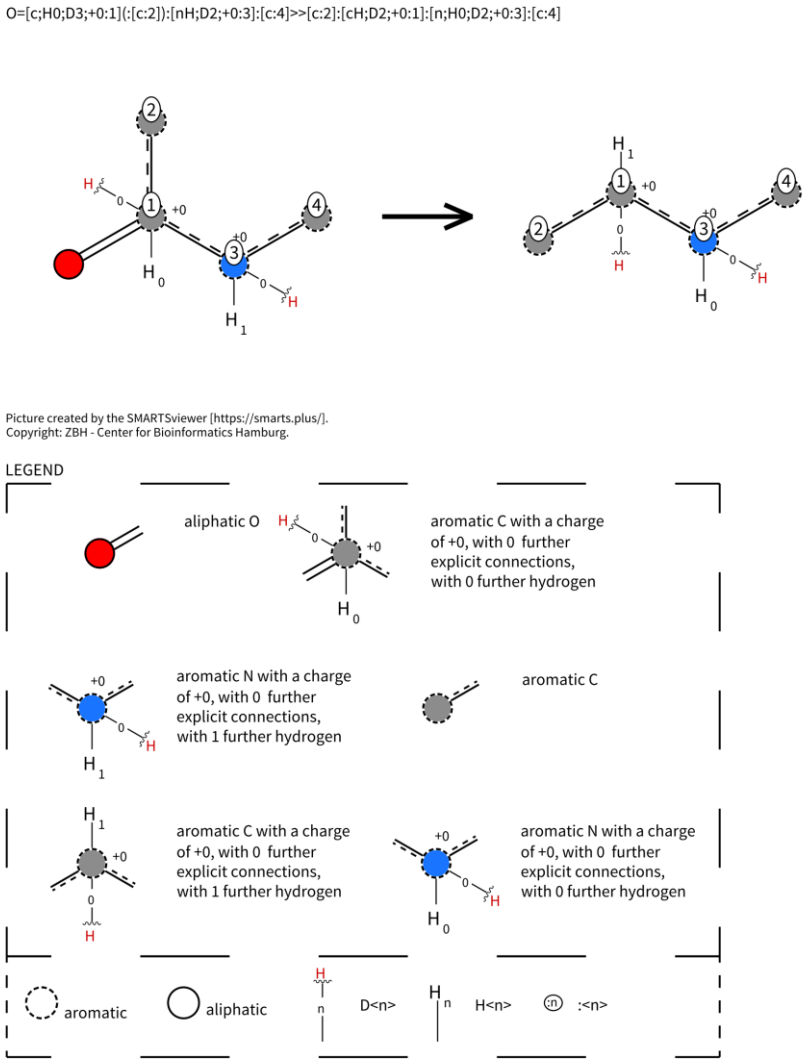
- Compound Identifiers
- Building Block Identifiers

In stock status

- Update in stock status of molecules

Reaction & Molecule Templates





SMARTSview - Visualize Chemical Patterns
https://smarts.plus/smartsview/46a141ab5f9a7485b32068764a54113a51ec970f72978b99bb523dee6917

Template Enrichment Pipeline

Process

- Generate Reaction Templates
  - Binary Reactions
  - Radius range

Store

- Store Reaction, Molecule Template nodes & Relationships

(1) Kannas, C.; Thakkar, A.; Bjerrum, E.; Genheden, S. Rxnutils – A Cheminformatics Python Library for Manipulating Chemical Reaction Data. 2022. https://doi.org/10.26434/chemrxiv-2022-wt440-v2.
(2) Thakkar, A.; Kogej, T.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. Datasets and Their Influence on the Development of Computer Assisted Synthesis Planning Tools in the Pharmaceutical Domain. Chem. Sci. 2019, 11 (1), 154–168. https://doi.org/10.1039/c9sc04944d.
(3) Coley, C. W.; Green, W. H.; Jensen, K. F. RDChiral: An RDKit Wrapper for Handling Stereochemistry in Retrosynthetic Template Extraction and Application. J. Chem. Inf. Model. 2019, 59 (6), 2529–2537. https://doi.org/10.1021/acs.jcim.9b00286.

Link Prediction - Reaction Template

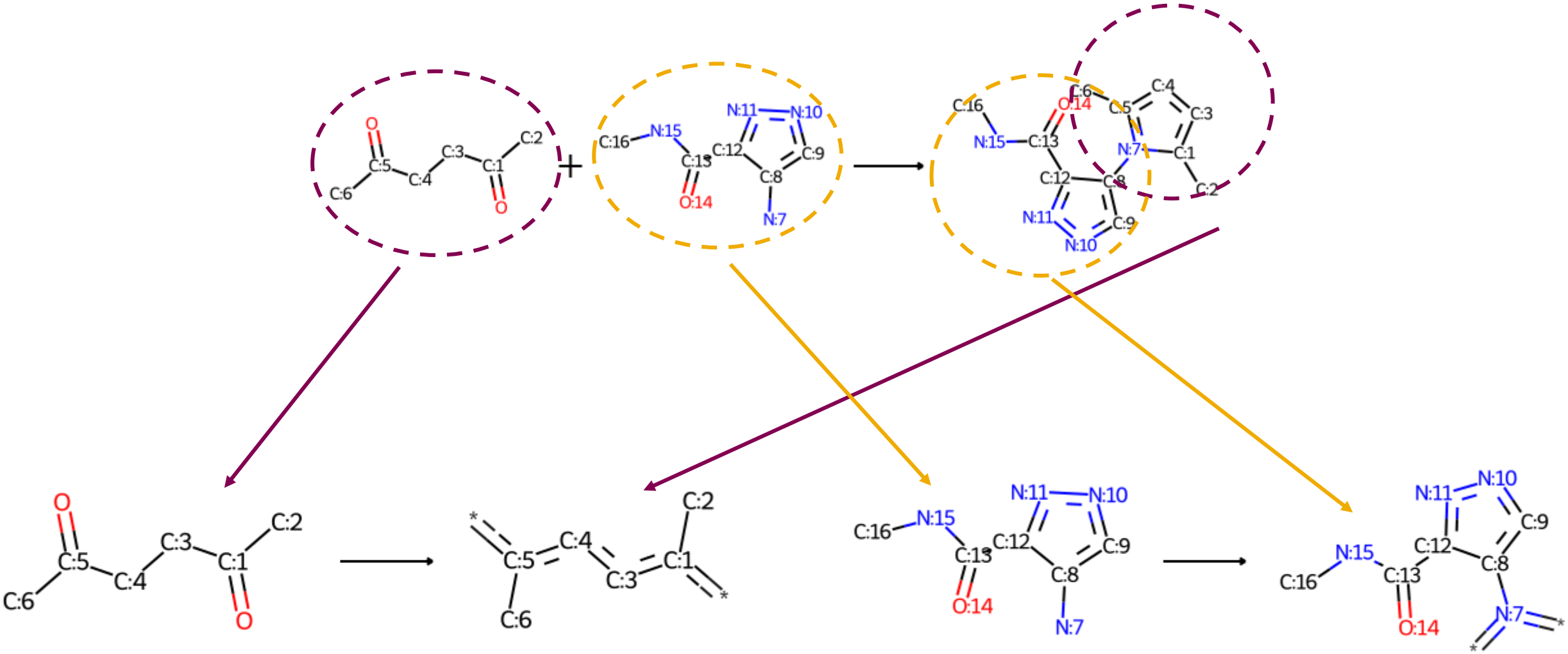
Process

- From Reaction Templates (Binary Reactions)
  - Generate Half-Templates

Store

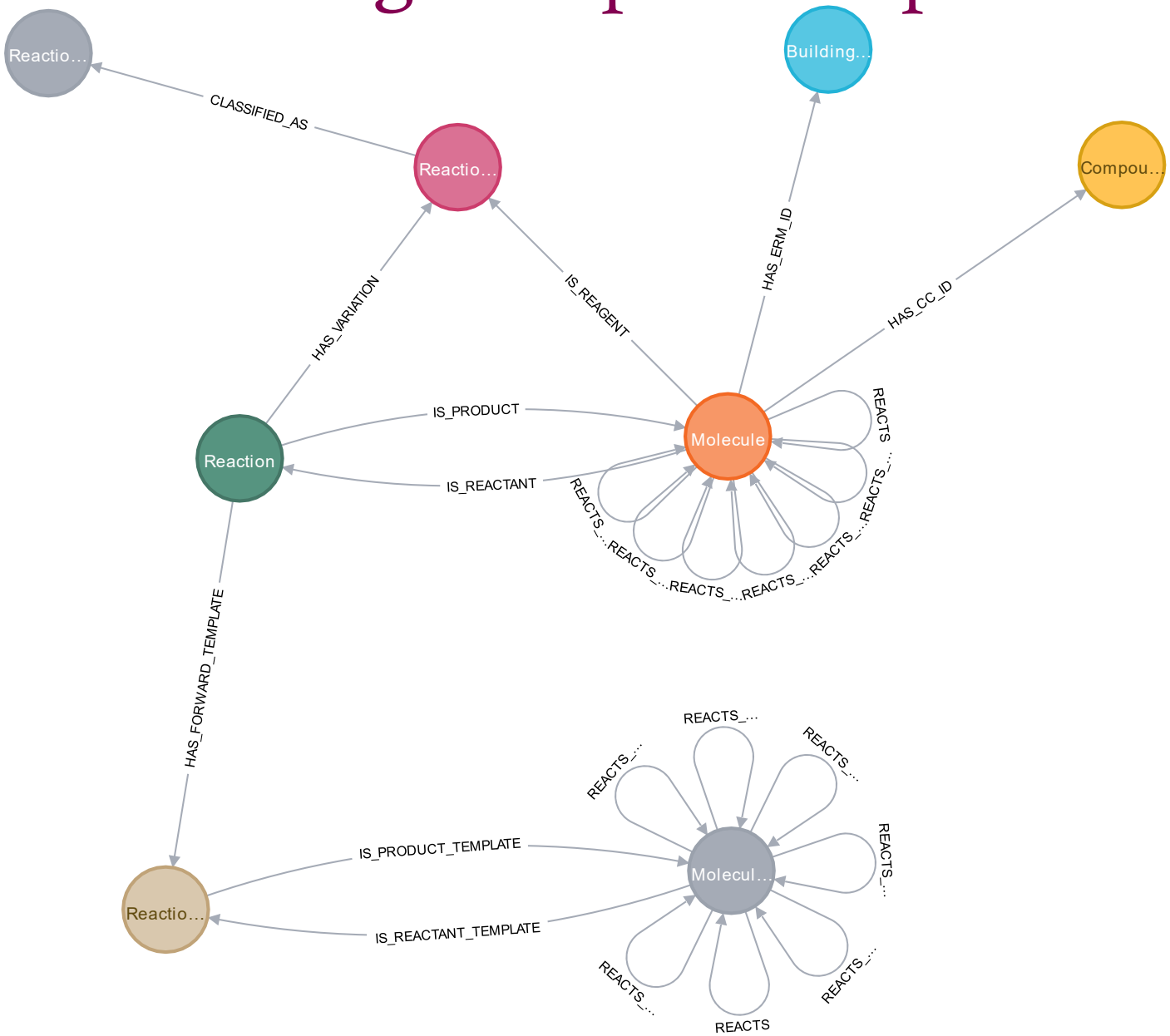
- Update Reaction Template nodes with Half-Templates

Half-Templates



Reaction template extracted from 1 USPTO reaction (17 reaction variations)

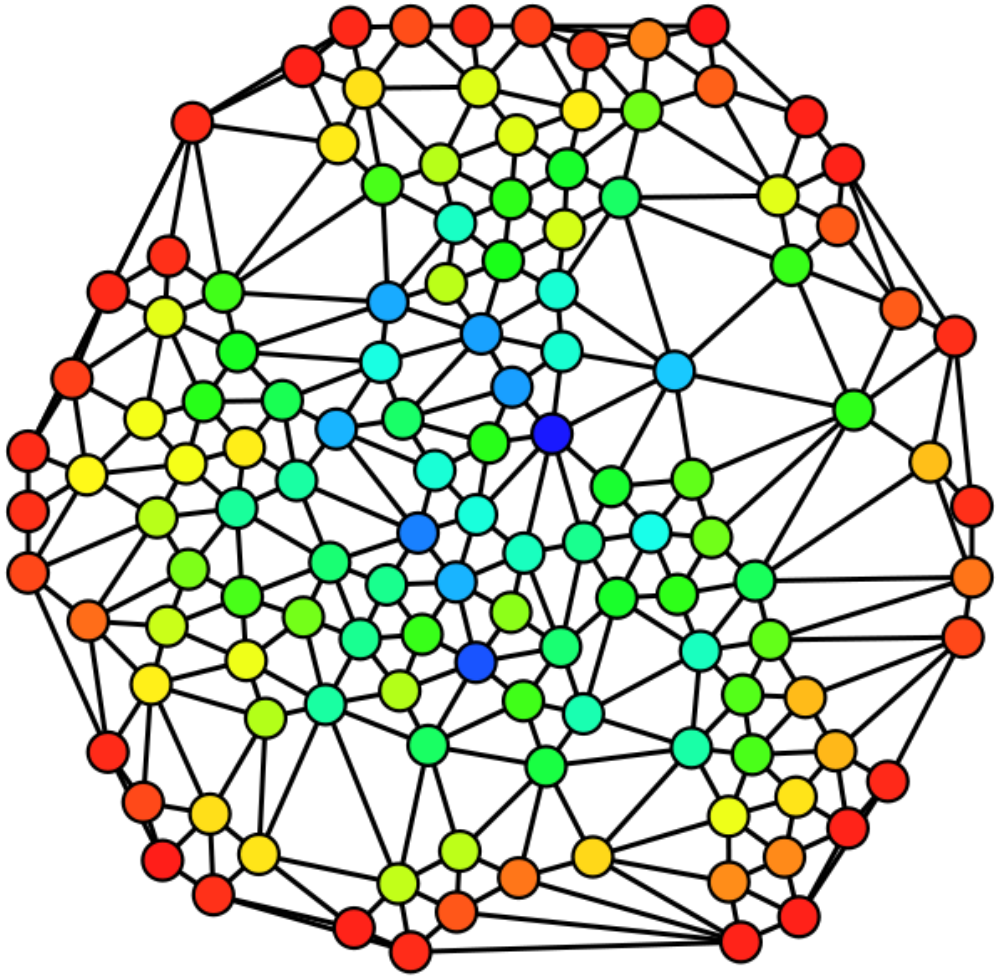
Reaction Knowledge Graph – Graph Model



3

Reaction Knowledge Graph – Using the graph

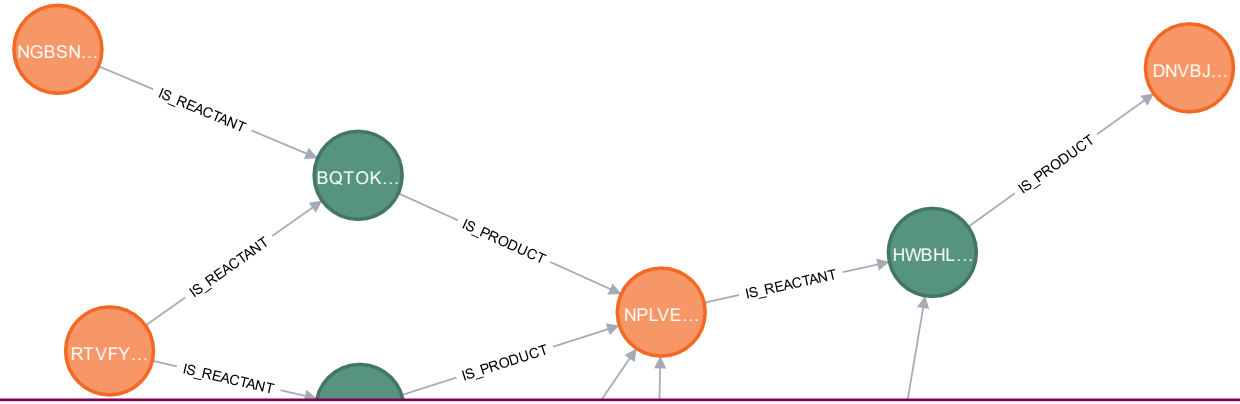
Graph Analytics



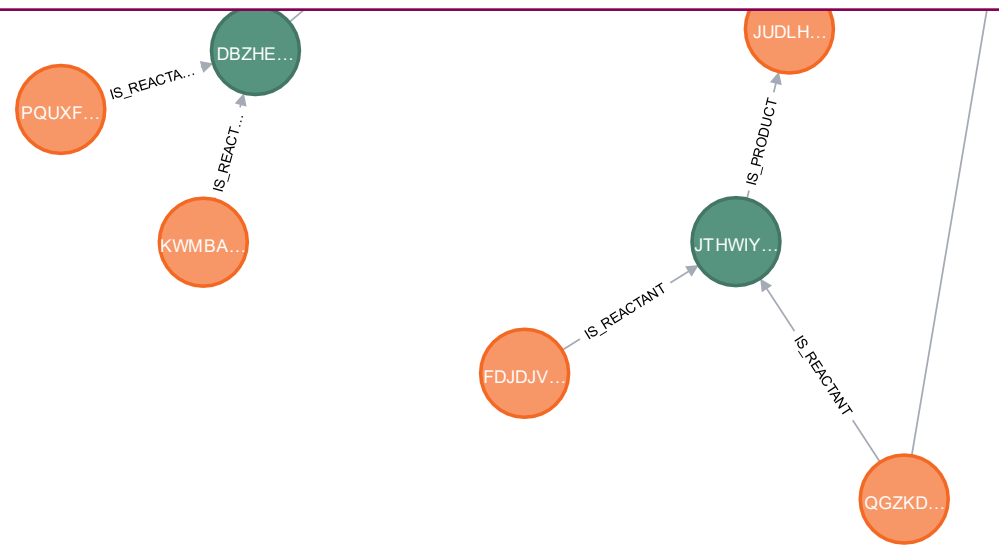
An undirected graph colored based on the betweenness centrality of each vertex from least (red) to greatest (blue).

- Important:
  - Reactions (maximize accessible chemical space)
  - Molecules (most needed chemicals)

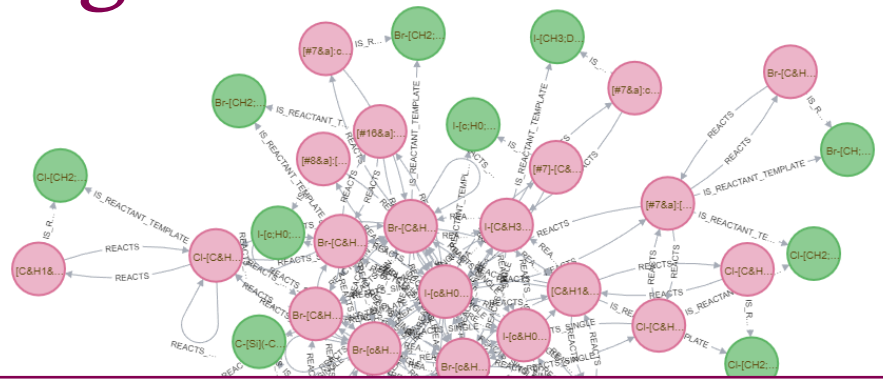
Synthesis Trees



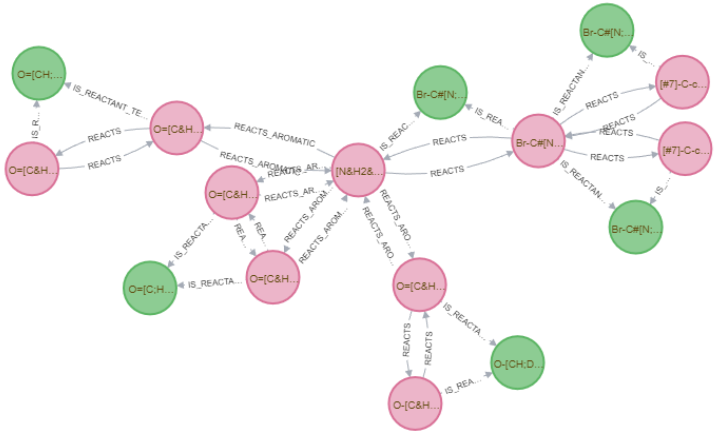

```
MATCH p=(e:Molecule)<-[:IS_PRODUCT|IS_REACTANT*2..10]-(s:Molecule)
WHERE s <> e
AND e.inchikey = $inchikey
WITH p, nodes(p) as path_nodes, size(relationships(p)) as path_length
WHERE path_length <= $max_depth AND size(path_nodes) =
size(apoc.coll.toSet(path_nodes))
RETURN p, size(apoc.coll.toSet(path_nodes)) as num_nodes, path_length
ORDER BY path_length asc
LIMIT $limit_val
```



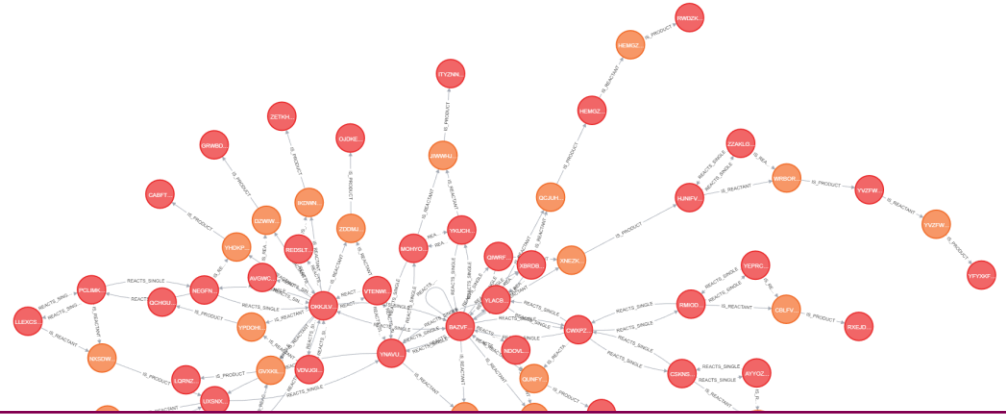
Link Prediction using Molecule & Reaction Templates



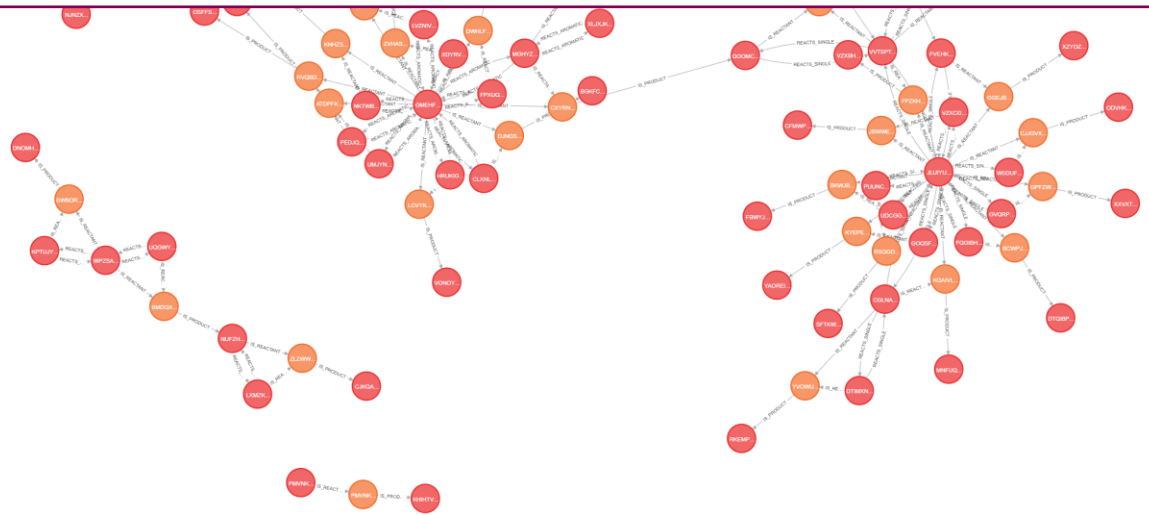
```
MATCH (RV:ReactionVariation {source_name: 'USPTO'}),
(R:Reaction)-[:HAS_VARIATION]->(RV),
(RT:ReactionTemplate {radius_1:True})<-[:HAS_FORWARD_TEMPLATE]-(R),
(RT)<-[IRT1:IS_REACTANT_TEMPLATE]-(MT1:MoleculeTemplate {radius_1:True})-[reacts:REACTS]-
(MT2:MoleculeTemplate {radius_1:True})-[IRT2:IS_REACTANT_TEMPLATE]->(RT)
WHERE MT1 <> MT2
RETURN RT, IRT1, MT1, reacts, MT2, IRT2
```



Link Prediction on Molecules & Reactions



```
MATCH (rv:ReactionVariation {source_name: 'USPTO'})-->(rc:ReactionClassification)
WHERE NOT rc.classification_id STARTS WITH '0'
WITH rv
MATCH (r:Reaction)-->(rv)
MATCH (rm:Molecule)-[rr:IS_REACTANT]->(r)-[pr:IS_PRODUCT]->(pm:Molecule)
RETURN rm, rr, r, pr, pm
```

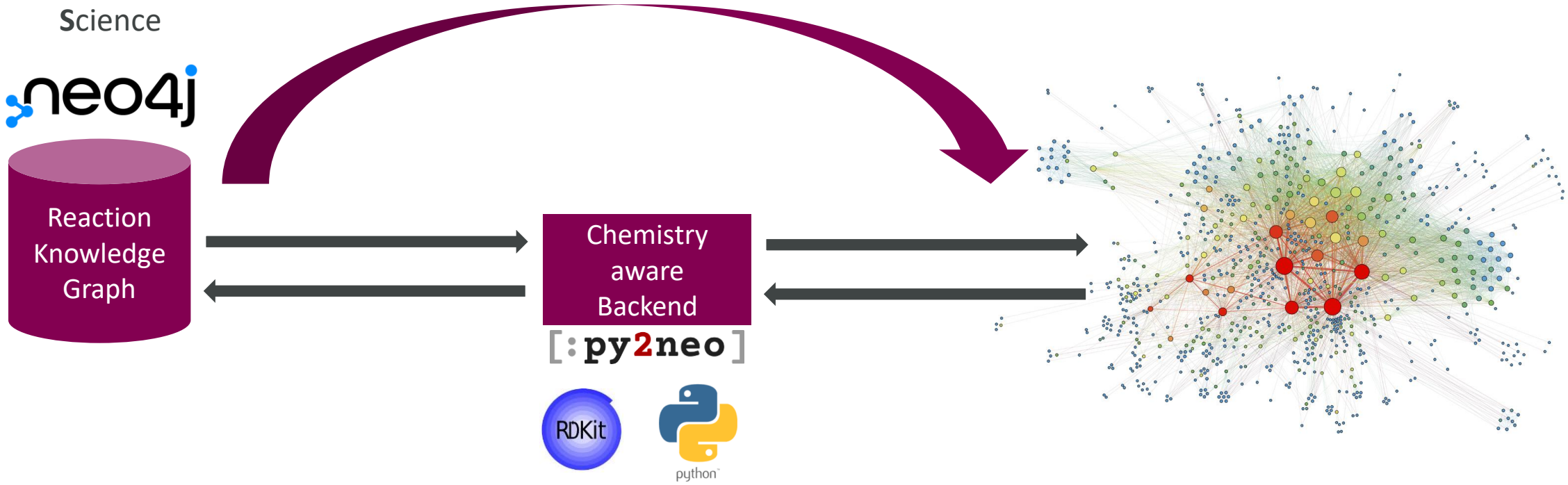


4

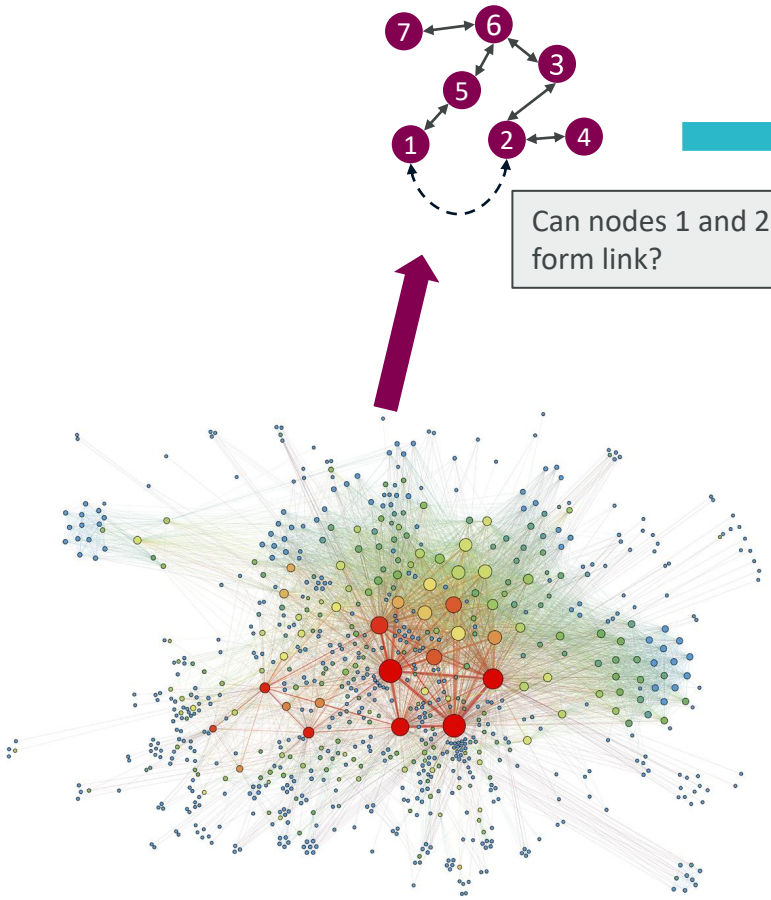
Future Work

Analytics on full AZ Collection

Graph
Data
Science



Link Prediction Workflows



Can nodes 1 and 2 form link?

ML for Link Prediction

Probability *p* of link

If *p* above cut-off:

1 + 2

Molecule Template Space

Product template construction

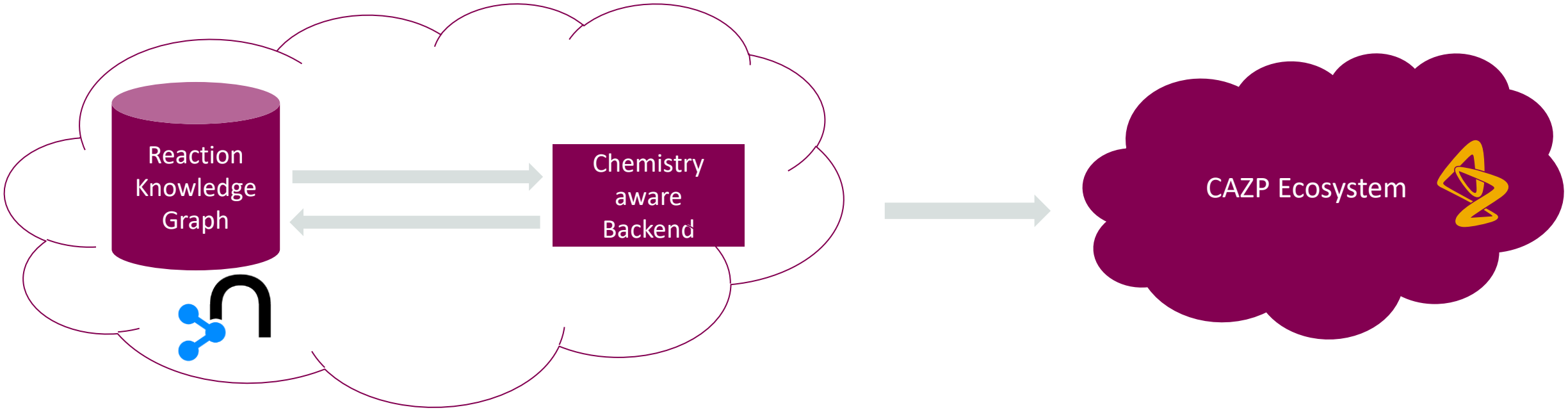
Transformer for Product Prediction

T^P

P

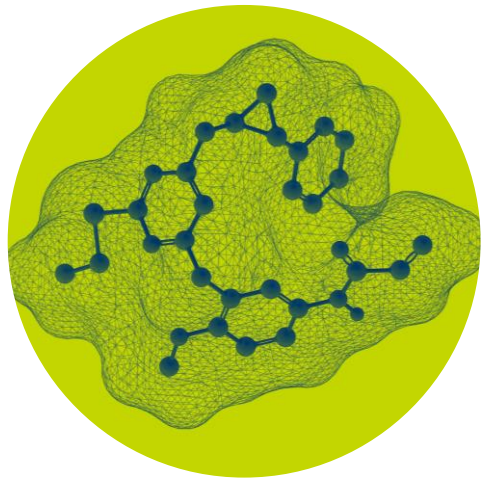
Molecule Space

Reaction Knowledge Graph – Computer-Aided Zynthesis Prediction Ecosystem Integration



Explore Chemical Reaction Space

Exploit Chemical Reaction Space



Assist Synthesis & Reaction Prediction

Open Source Software

Neo4j CE

Py2Neo

RDKit

[rxnutils]

Editable MolObjects

RDChiral

[rxnutils]

Pseudo standard templates

> Template Fingerprints <

Acknowledgements

Molecular AI

- Tomas Bastys
- Emma Rydholm
- Emma Svensson
- Thierry Kogej
- Samuel Genheden
- Ola Engkvist

R&D IT

- AJ Vignesh
- Rathi Prakash
- Alla Bushoy

Thank you.

Molecular AI Open Positions

- AI Scientist – AI and Machine Learning Expert
- Computational Scientist – Molecular Dynamics Expert with Interest in Machine Learning
- Machine learning engineer with chemistry applications

Deadline: 16th Oct 2022