Vernalis
RESEARCH
part of Vernalis plc

Exploration of Patent Chemistry by Fuzzy MCS-led Fragment Decomposition

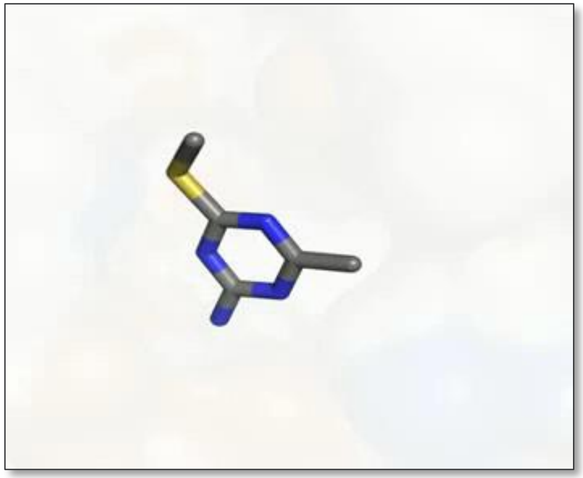
Richard Sherhod

r.sherhod@vernalis.com

THE QUEEN'S AWARDS
FOR ENTERPRISE:
INTERNATIONAL TRADE
2014

About Vernalis

- Expertise
  - Fragments and structure-based drug discovery (Protein Science, Structural Biology, Chemistry)
- Therapeutic areas
  - Oncology, CNS, infectious diseases
- Location
  - Based in Granta Park, outside Cambridge, UK



Cambridge
GrantaPark
M1
A1M
Stansted
M11
M1
M25
M40
M25
London
M4
Heathrow
M2
M3
M25
M20
Gatwick
M23

Contents

- Our aims
  - Our solution
- Patent data-mining
  - SAR table extraction
  - Data curation
- Fuzzy MCS analysis
  - Maximum common substructure mining
  - Fuzzy MCS
  - Recursive MCS mining
  - Scaffold-tree formation
- Fragment decomposition
  - MCS-led decomposition
  - Implementation
- Conclusion

Our Aims

Our Aims

The situation:

- New project with existing target-related patents
- …or existing project with newly-published target-related patents
- We need to understand the chemical space covered by the patents

Our aims:

- To summarise the coverage of exemplified structures
- To extract and present relevant SAR data

Our Aims

Our Solution

- KNIME based application for exploring patent chemistry
  - Accessed via the KNIME Web Portal

- KNIME workflows for preparing data and presenting results
  - Patent processing workflows (admin):
    - Structures and data extraction and curation
    - Recursive fuzzy MCS mining
  - Interactive results workflow
    - MCS tree formation for visualisation
    - Fragment decomposition for chosen MCS

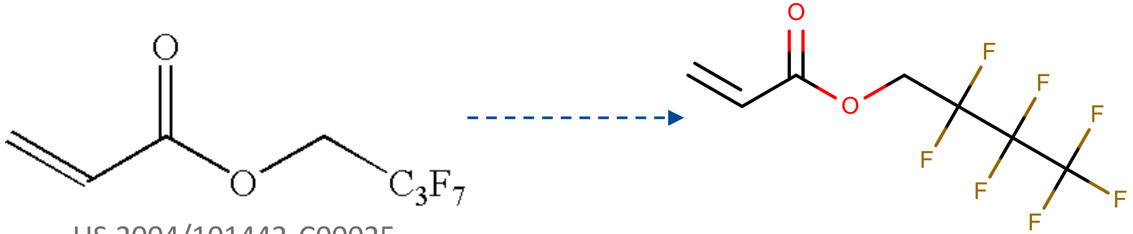
Vernalis
RESEARCH
part of Vernalis plc

Patent Data Mining

THE QUEEN'S AWARDS
FOR ENTERPRISE:
INTERNATIONAL TRADE
2014

Patent Data Mining

- Patent data mined with NextMove Software's LeadMine
  - Processes HTML, XML, raw text etc.
  - Automatic structure extraction
    - Text-to-structure – IUPAC, generic names, abbreviations
    - CDX-to-structure – Ambiguities, drawing errors e.g. floating alkanes



US 2004/101442 C00025

May, J., Lowe, D. & Sayle, R., 2016. Sketchy Sketches: Hiding Chemistry in Plain Sight. Seventh Joint Sheffield Conference on Chemoinformatics. Available at: http://cisrg.shef.ac.uk/shef2016/talks/poster21.pdf [Accessed September 14, 2017].

- Patents accessed and processed by PatFetch web service
  - Patent archives stored locally

Patent Data Mining

SAR Table Extraction

- Structures extracted from:
  - Names
  - CDX images
  - R-group tables
- Table data – activity, properties etc.
  - New feature of LeadMine
- IDs and data associated with structures

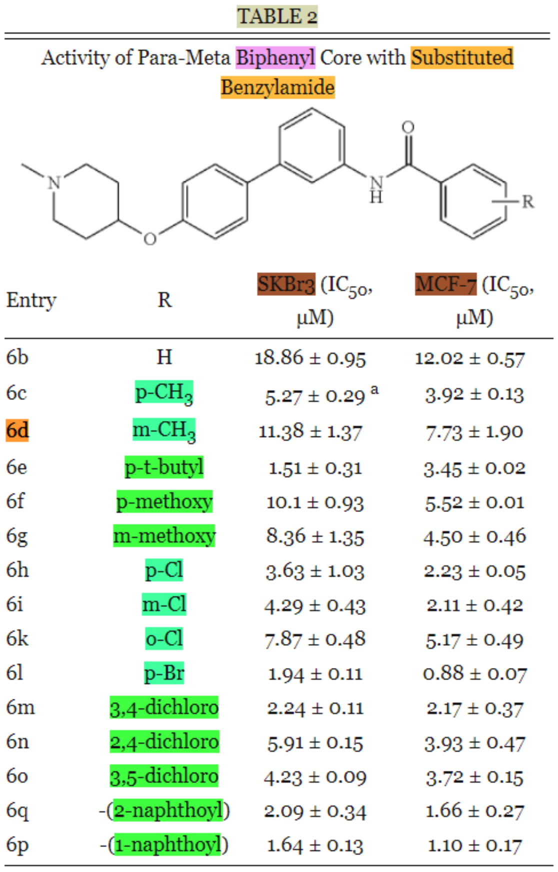
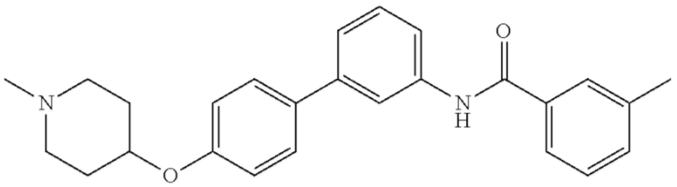


TABLE 2

Activity of Para-Meta Biphenyl Core with Substituted Benzylamide

Entry	R	SKBr3 (IC_{50}, μM)	MCF-7 (IC_{50}, μM)
6b	H	18.86 ± 0.95	12.02 ± 0.57
6c	p-CH_3	5.27 ± 0.29 [a]	3.92 ± 0.13
6d	m-CH_3	11.38 ± 1.37	7.73 ± 1.90
6e	p-t-butyl	1.51 ± 0.31	3.45 ± 0.02
6f	p-methoxy	10.1 ± 0.93	5.52 ± 0.01
6g	m-methoxy	8.36 ± 1.35	4.50 ± 0.46
6h	p-Cl	3.63 ± 1.03	2.23 ± 0.05
6i	m-Cl	4.29 ± 0.43	2.11 ± 0.42
6k	o-Cl	7.87 ± 0.48	5.17 ± 0.49
6l	p-Br	1.94 ± 0.11	0.88 ± 0.07
6m	3,4-dichloro	2.24 ± 0.11	2.17 ± 0.37
6n	2,4-dichloro	5.91 ± 0.15	3.93 ± 0.47
6o	3,5-dichloro	4.23 ± 0.09	3.72 ± 0.15
6q	-(2-naphthoyl)	2.09 ± 0.34	1.66 ± 0.27
6p	-(1-naphthoyl)	1.64 ± 0.13	1.10 ± 0.17

US-20160272584-A1



6d

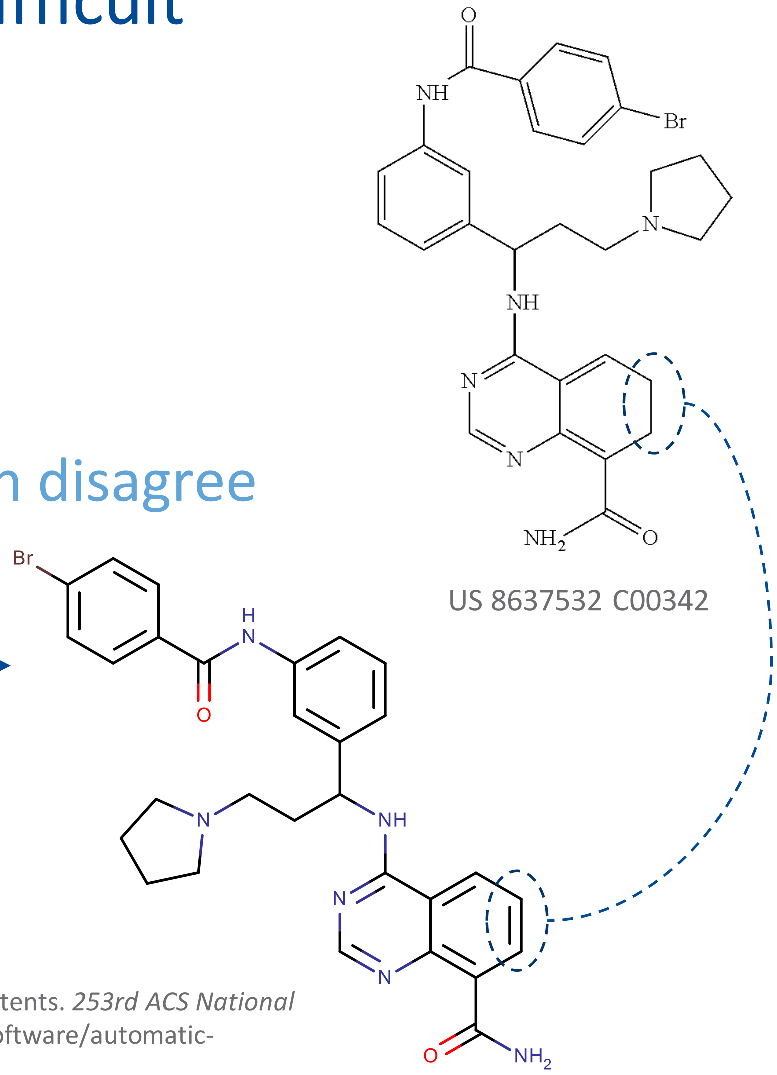
3-methyl-N-(4'-((1-methylpiperidin-4-yl)oxy)-[1,1'-biphenyl]-3-yl)benzamide (6d)

Patent Data Mining

Interactive Data Curation

- Extracting clean patent data is difficult
  - Patent tables can contain errors:
    - Typos
    - OCR errors
    - Missing values
    - Inconsistent labelling schemes
  - Structure names and CDX images can disagree

4-{1-[3-(4-Bromo-benzoyl-amino)-phenyl]-3-pyrrolidin-1-yl-propyl-amino}-quinazoline-8-carboxylic acid amide



US 8637532 C00342

  - No two patents are alike

Lowe, D., Senger, S. & Sayle, Ro., 2017. Automatic extraction of bioactivity data from patents. *253rd ACS National Meeting, San Francisco, CA, USA*. Available at: https://www.slideshare.net/NextMoveSoftware/automatic-extraction-of-bioactivity-data-from-patents-74402139 [Accessed September 14, 2017].

Patent Data Mining

Interactive Data Curation

- Fully automated extraction is unattainable (for now)
  - Structures and data are interactively assessed
  - Data may be exported for manual editing
  - Modified data is semi-automatically validated

Vernalis
RESEARCH
part of Vernalis plc

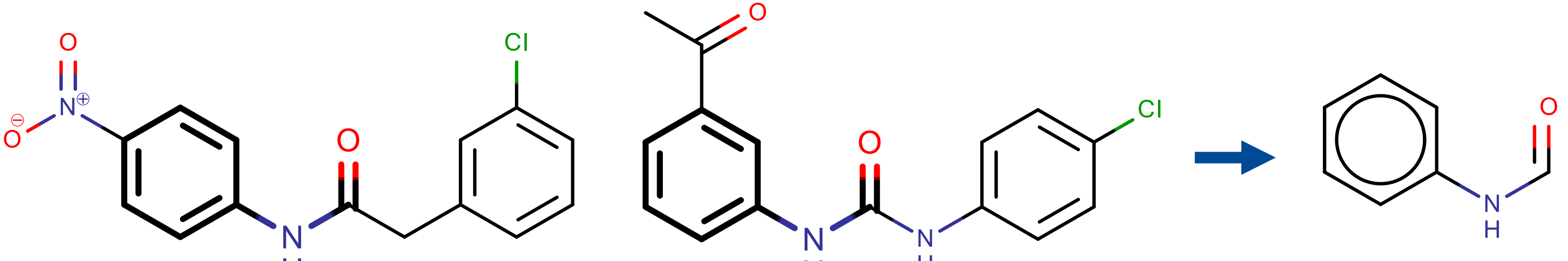
Fuzzy MCS Analysis

THE QUEEN'S AWARDS
FOR ENTERPRISE:
INTERNATIONAL TRADE
2014

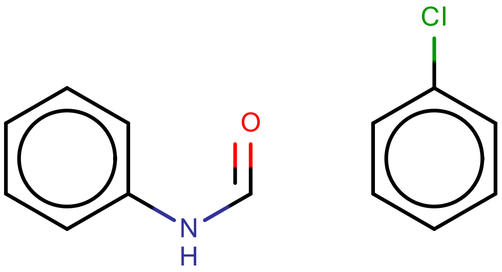
Fuzzy MCS Analysis

Maximum Common Substructure Mining

- The largest substructure common to a set of structures



  - …or disconnected substructures

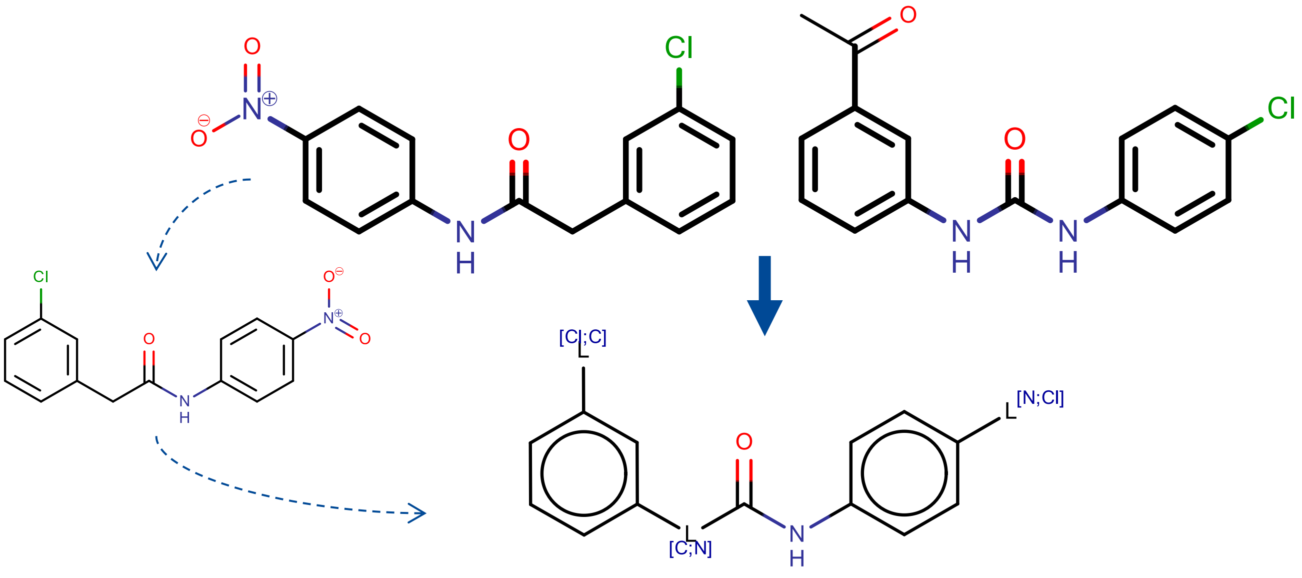


- Many implementations and uses in cheminformatics
- Traditionally only exact atom and bond matches allowed

Fuzzy MCS Analysis

Fuzzy MCS

- MCS with variation in atoms and/or bonds



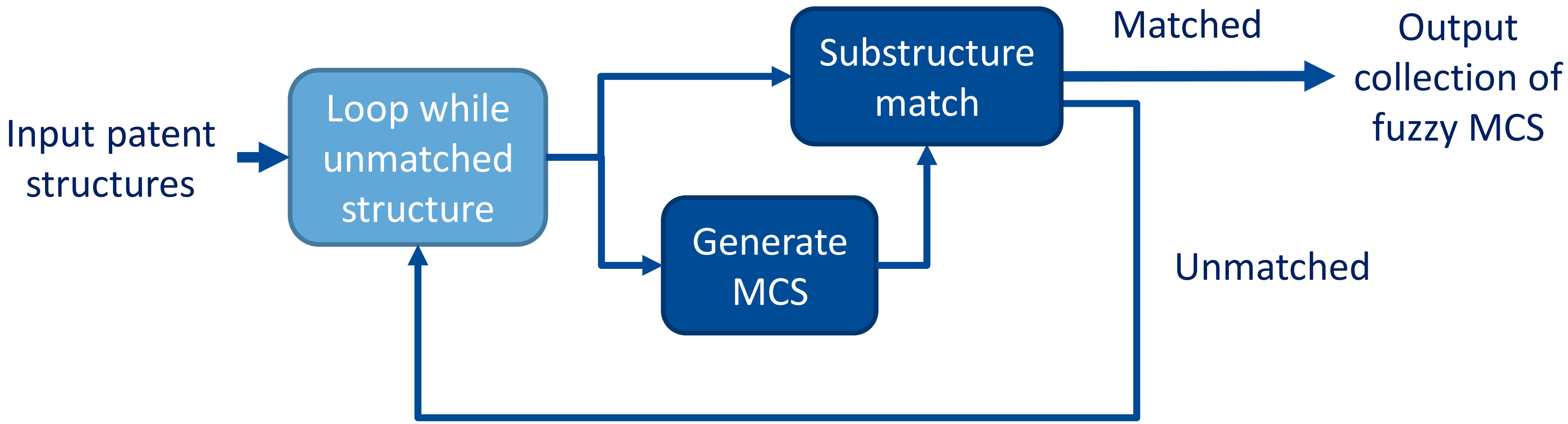
Fuzzy MCS

- Algorithm available in RDKit
  - Contributed by Andrew Dalke in 2012
  - RDKit MCS KNIME node

Fuzzy MCS Analysis

Recursive Fuzzy MCS Mining

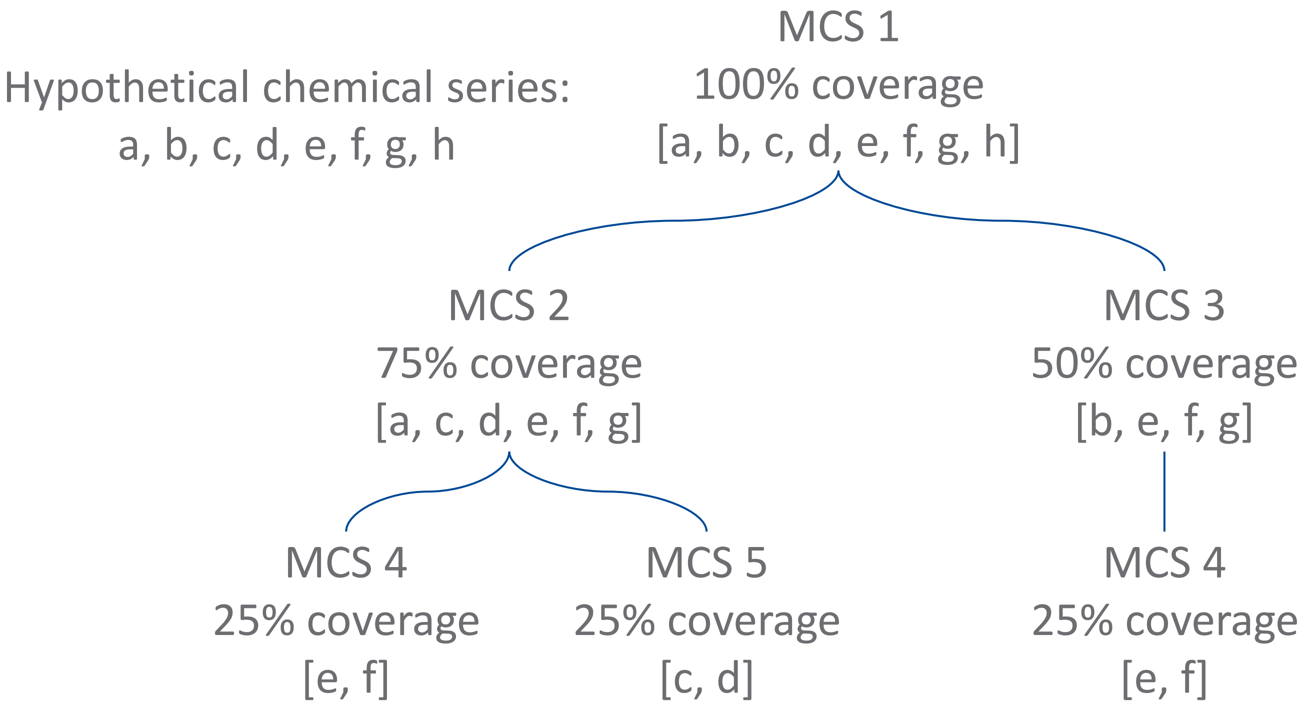
- Set of fuzzy MCS generated from patent structures
  - MCS generated for a range of coverage thresholds
    - i.e. MCS that represent 10%, 20%, 30%... 100% of input structures
  - Structures not covered by MCS are used to generate new MCS



Fuzzy MCS Analysis

MCS Tree Formation

- Resulting collection of MCS arranged into trees
  - Hierarchical relationship between MCS coverage
  - Fuzzy (overlapping) hierarchical clustering



Hypothetical chemical series:
a, b, c, d, e, f, g, h

MCS 1
100% coverage
[a, b, c, d, e, f, g, h]

MCS 2
75% coverage
[a, c, d, e, f, g]

MCS 3
50% coverage
[b, e, f, g]

MCS 4
25% coverage
[e, f]

MCS 5
25% coverage
[c, d]

MCS 4
25% coverage
[e, f]

Fuzzy MCS Analysis

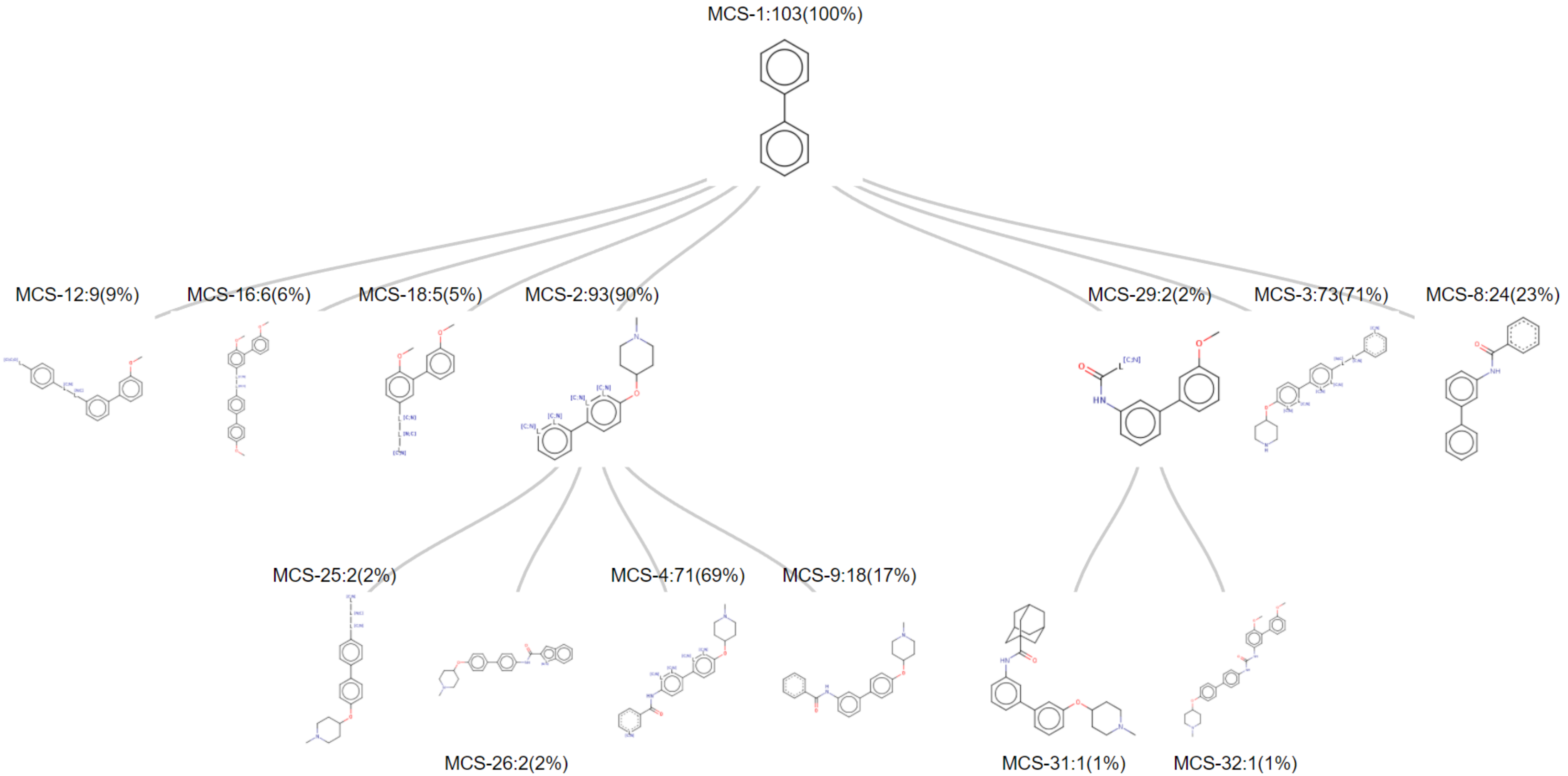
MCS Tree Formation - Visualisation

- Custom code used to generate and visualise tree structure
  - Tree represented as JSON object

- Tree presented as interactive view in KNIME Web Portal
  - Tree visualised in D3.js
  - Crude POC

Fuzzy MCS Analysis

MCS Tree Formation - Visualisation



Vernalis

RESEARCH

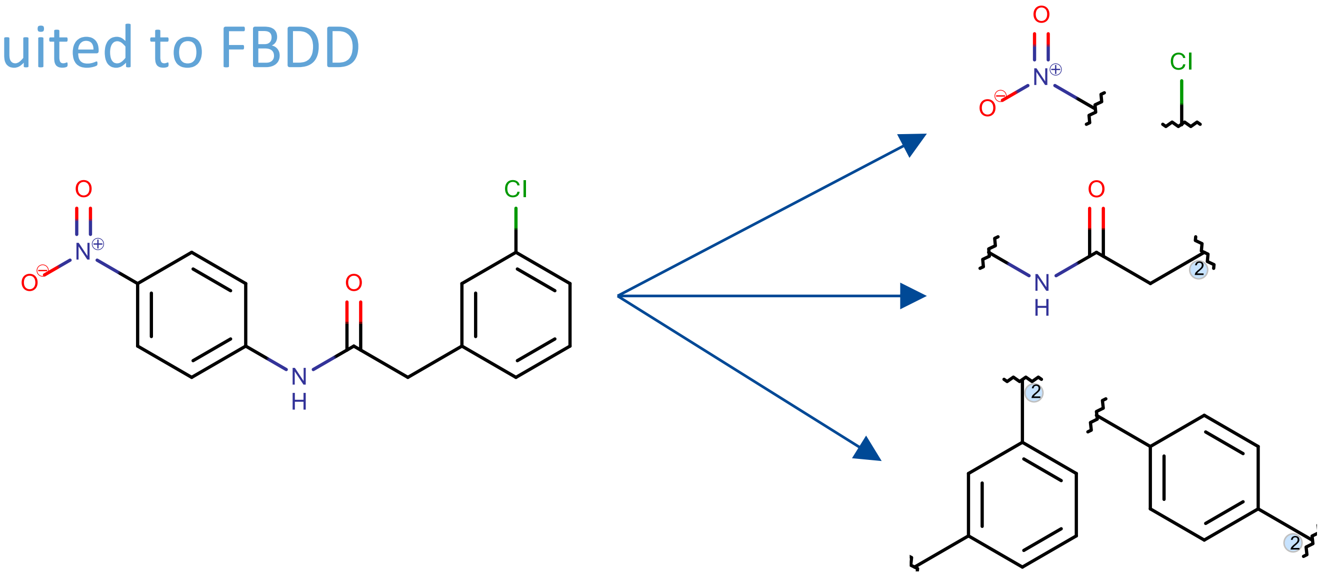
part of Vernalis plc

Fragment Decomposition

THE QUEEN'S AWARDS
FOR ENTERPRISE:
INTERNATIONAL TRADE
2014

Fragment Decomposition

- Break structures down into categorised fragments
  - R-groups
  - Chains – terminal or linkers
  - Rings – fused systems
- R-group decomposition
  - No scaffold
  - Suited to FBDD



R-groups

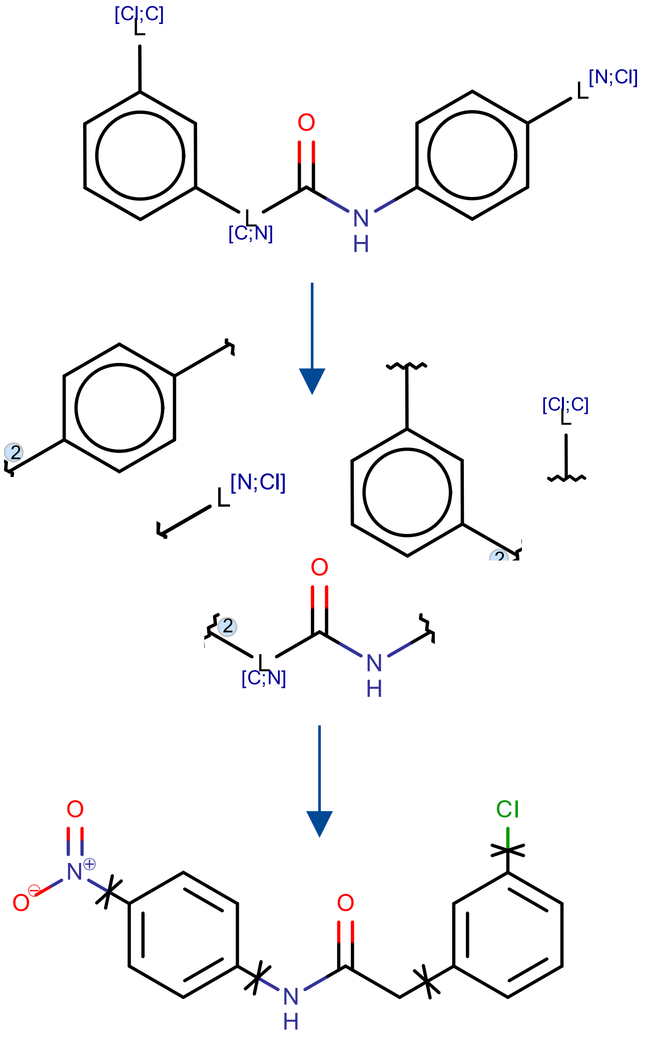
Chains

Rings

Fragment Decomposition

MCS-Led Decomposition

- Fuzzy MCS used to define fragment framework
- Fuzzy MCS decomposed into:
  - R-groups, Rings, Chains
  - Rings with variable features
  - Chains with variable features
- Fuzzy MCS fragments used to decompose structures



Fragment Decomposition

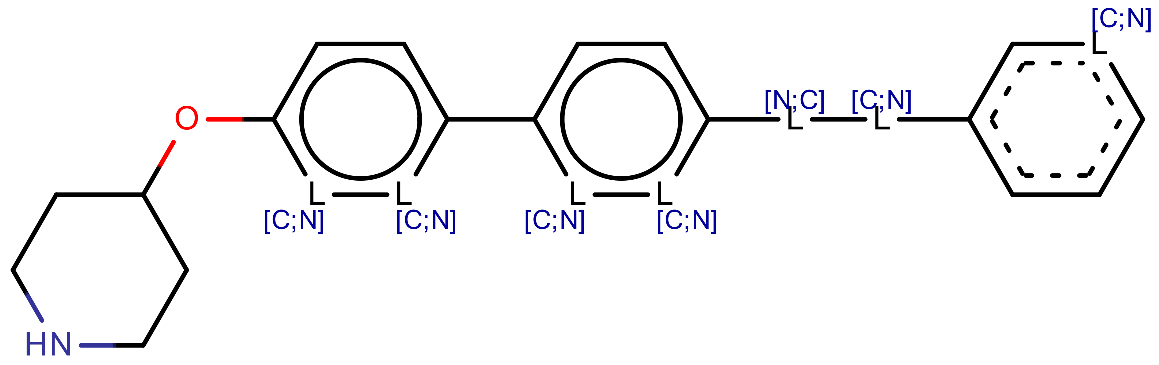
Implementation

- In-house algorithm developed with RDKit Java API

- Fragment Decomposition KNIME node released internally
  - Thanks Steve Roughley!

Fragment Decomposition

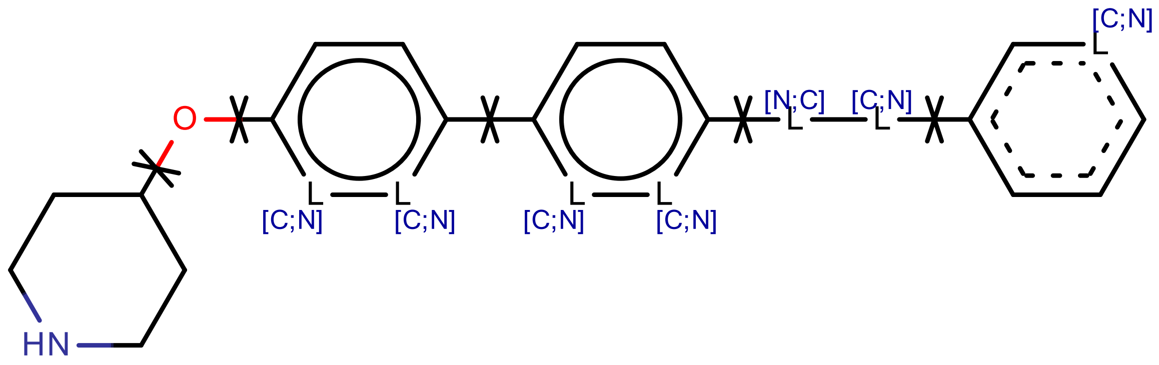
Implementation – Algorithm



- Fuzzy MCS is fragmented
  - Break all non-aromatic bonds between ring and non-ring atoms
    - SMARTS: **[!R0]!@&!:***

Fragment Decomposition

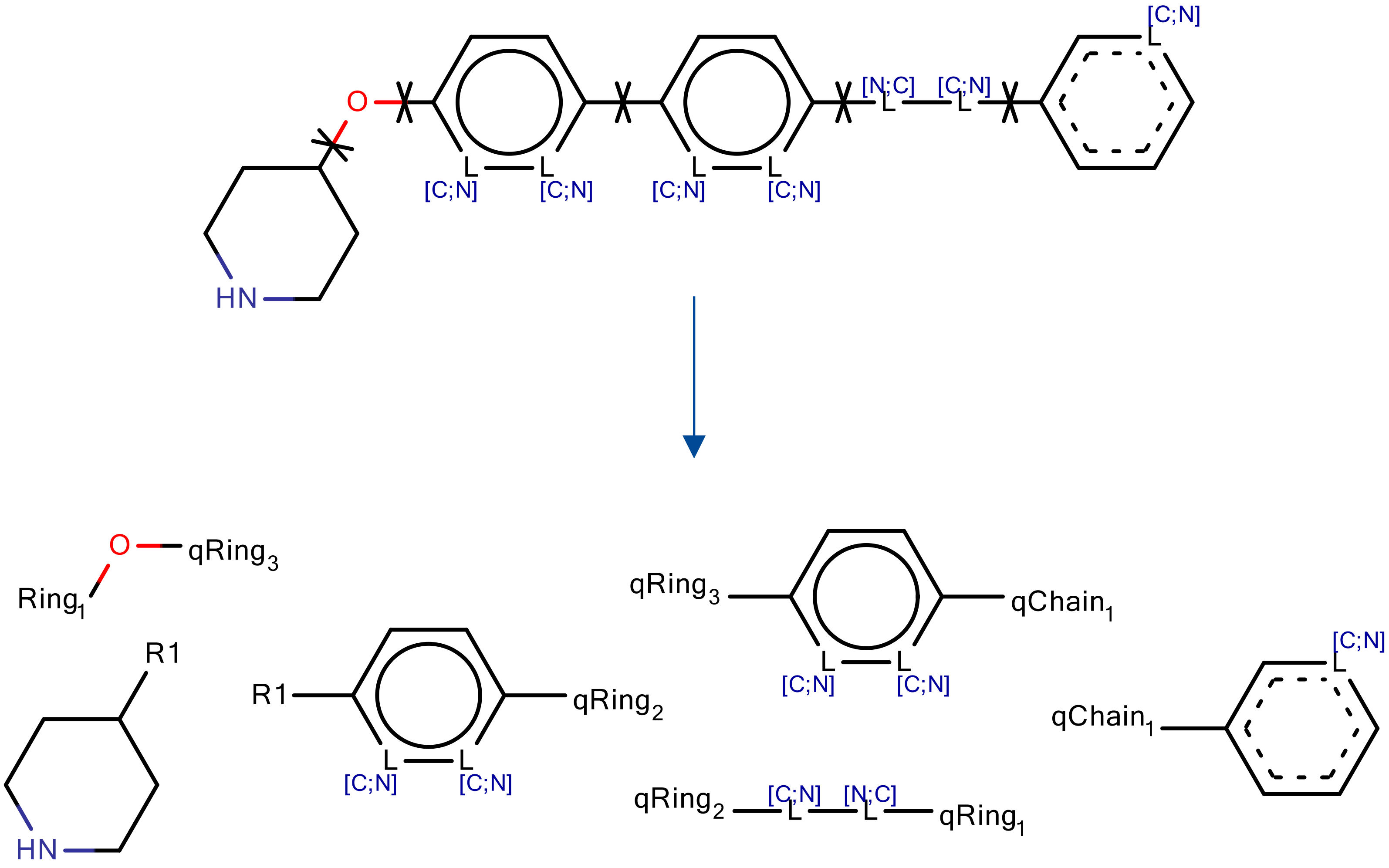
Implementation – Algorithm



- Fuzzy MCS is fragmented
  - Break all non-aromatic bonds between ring and non-ring atoms
    - SMARTS: **[!R0]!@&!:***
- Fragments are categorised
  - R-group, Ring, Chain, qRing, qChain
- Fragments given canonical identifiers

Fragment Decomposition

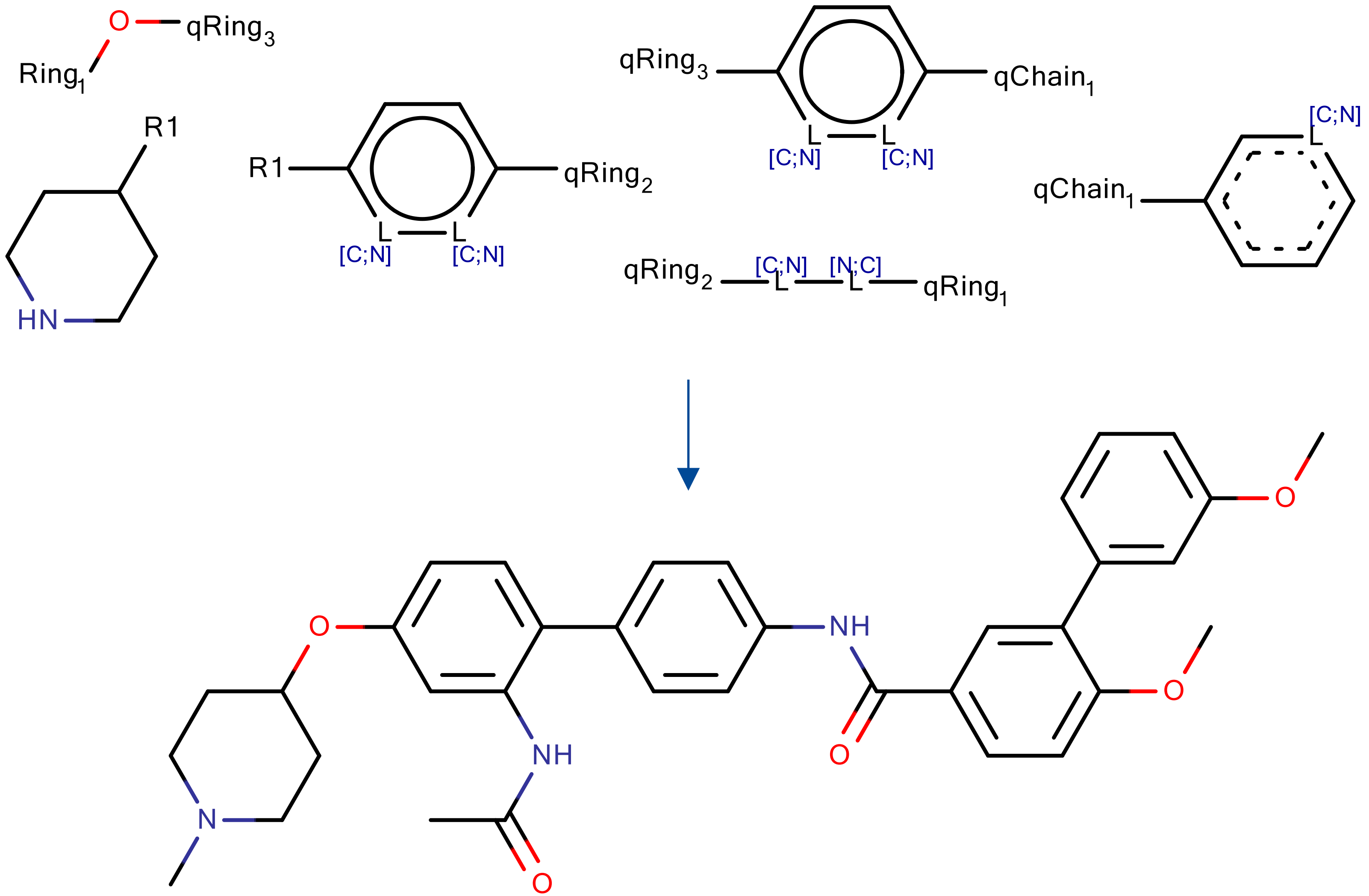
Implementation – Algorithm



Fragment Decomposition

Implementation – Algorithm

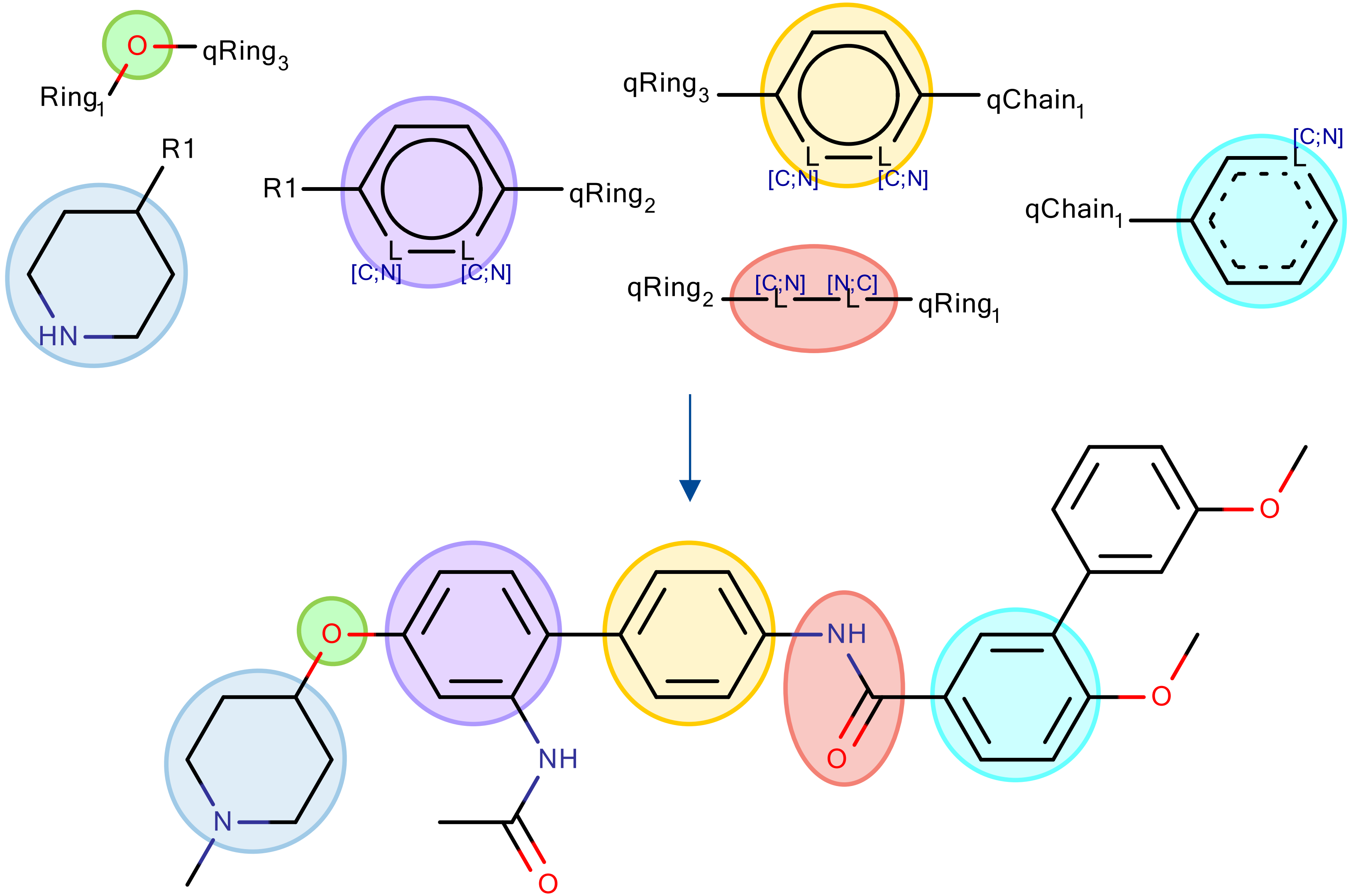
- Fragments are mapped to input structures



Fragment Decomposition

Implementation – Algorithm

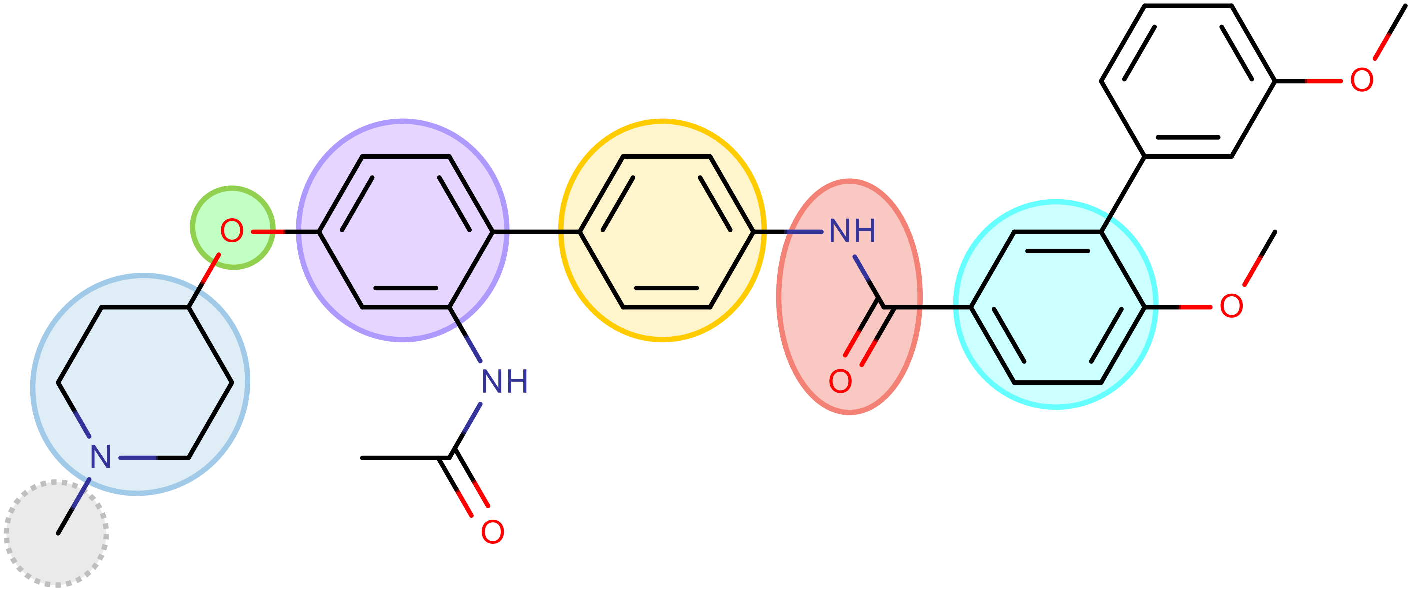
- Fragments are mapped to input structures



Fragment Decomposition

Implementation – Algorithm

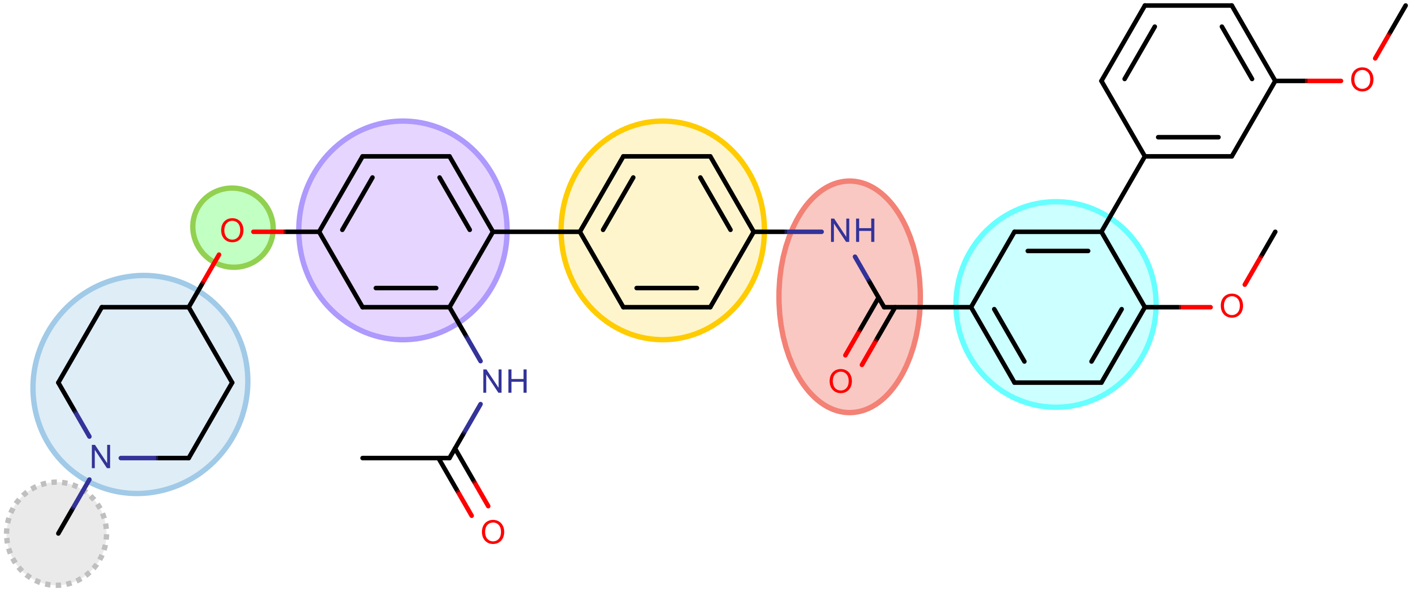
- Fused aromatic ring systems are expanded
  - Aromatic ring bonds are not broken
- Rings and chains are fragmented further
  - IDs of additional R-groups are canonicalised in a later process



Fragment Decomposition

Implementation – Algorithm

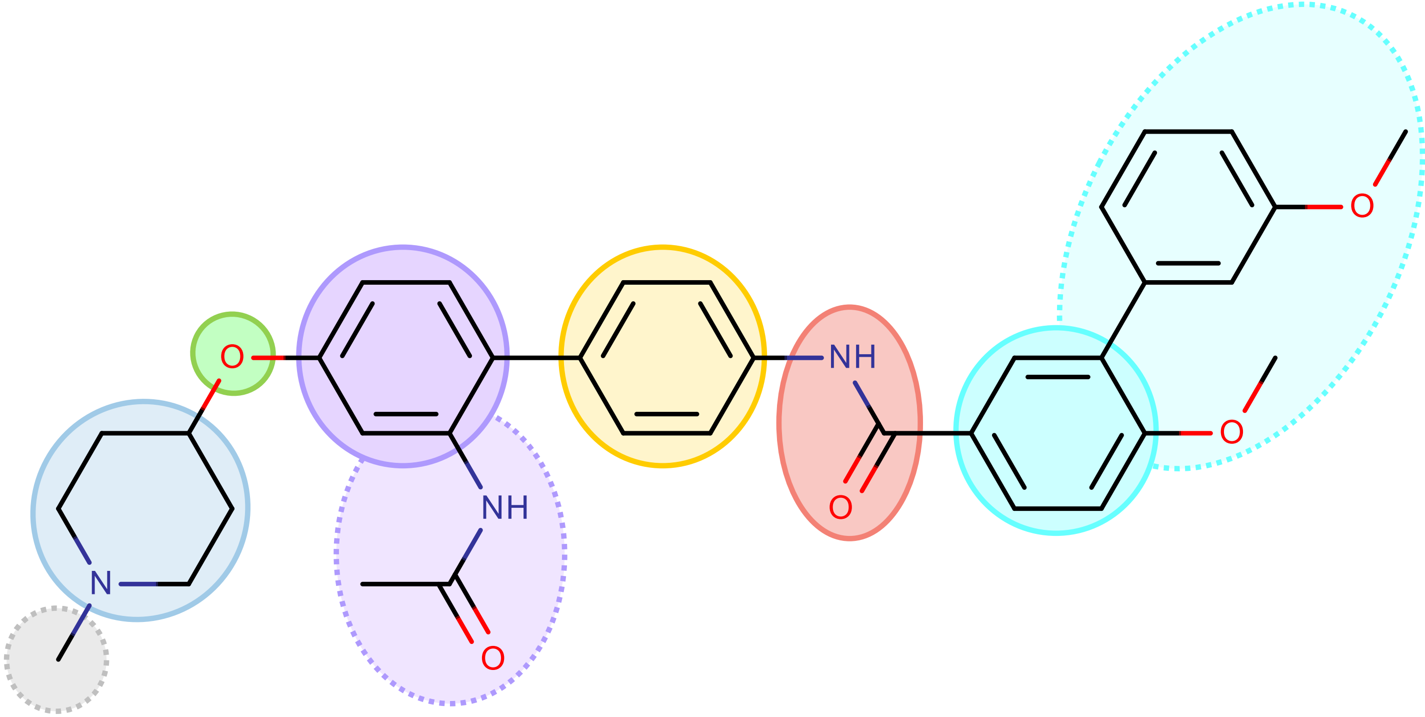
- qRings and qChains are either:



Fragment Decomposition

Implementation – Algorithm

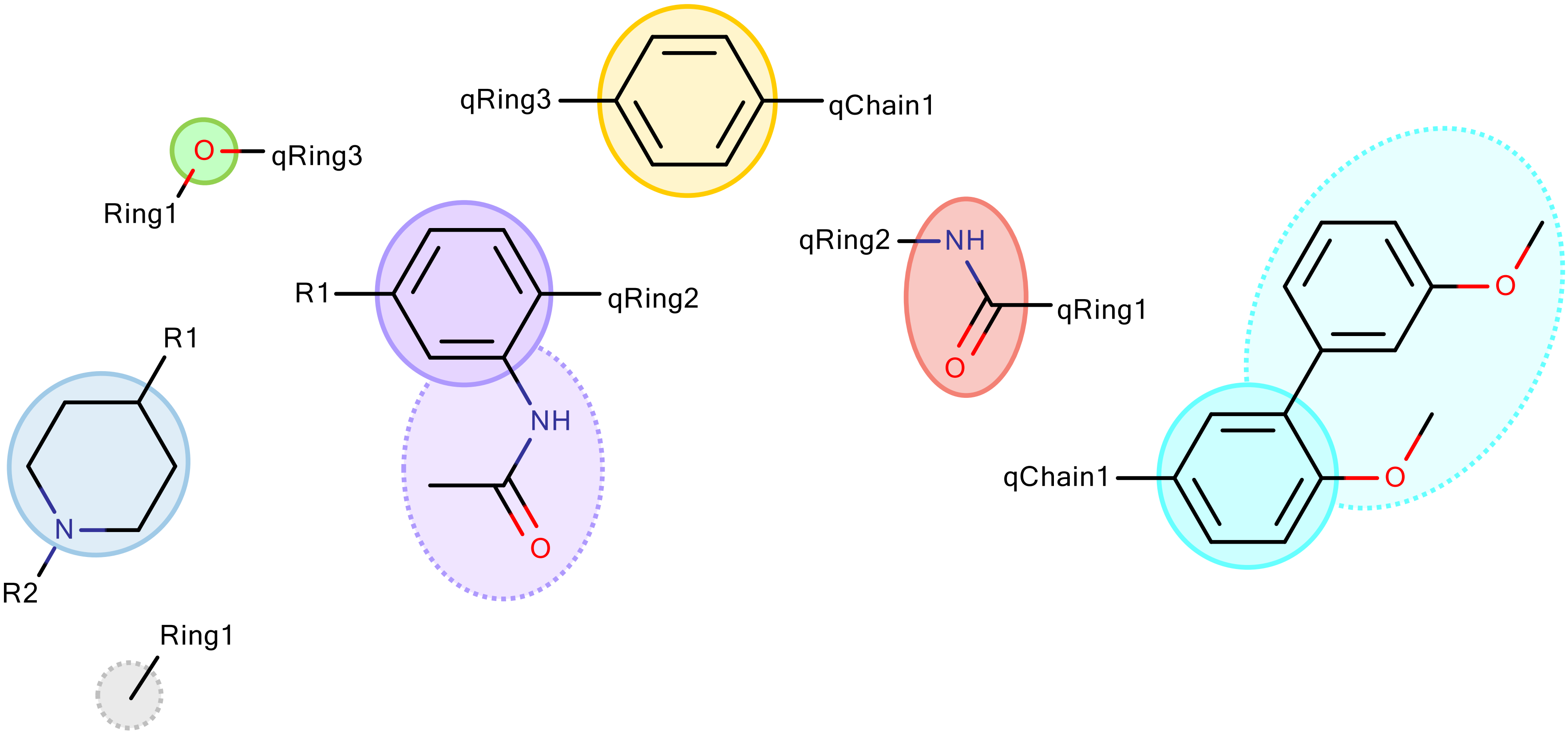
- qRings and qChains are either:
  - Expanded to incorporate additional features



Fragment Decomposition

Implementation – Algorithm

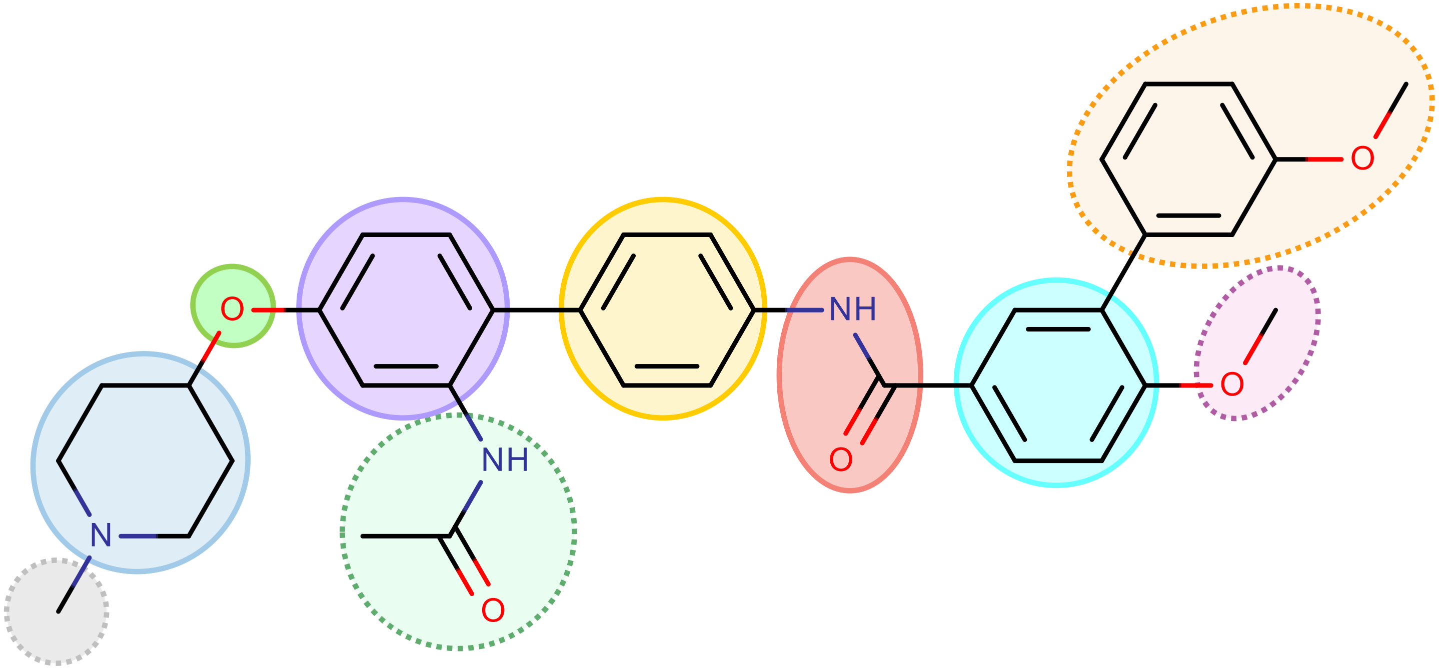
- qRings and qChains are either:
  - Expanded to incorporate additional features



Fragment Decomposition

Implementation – Algorithm

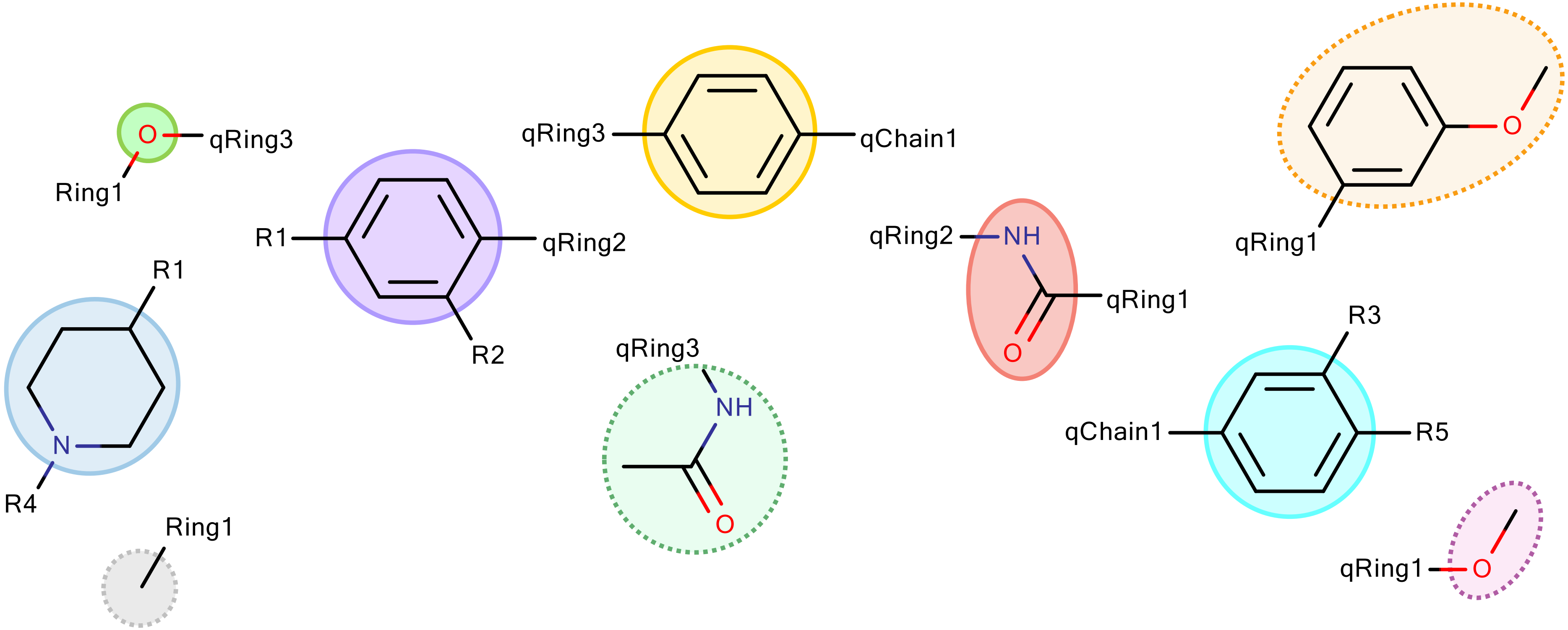
- qRings and qChains are either:
  - Expanded to incorporate additional features
  - Fragmented further into new R-groups



Fragment Decomposition

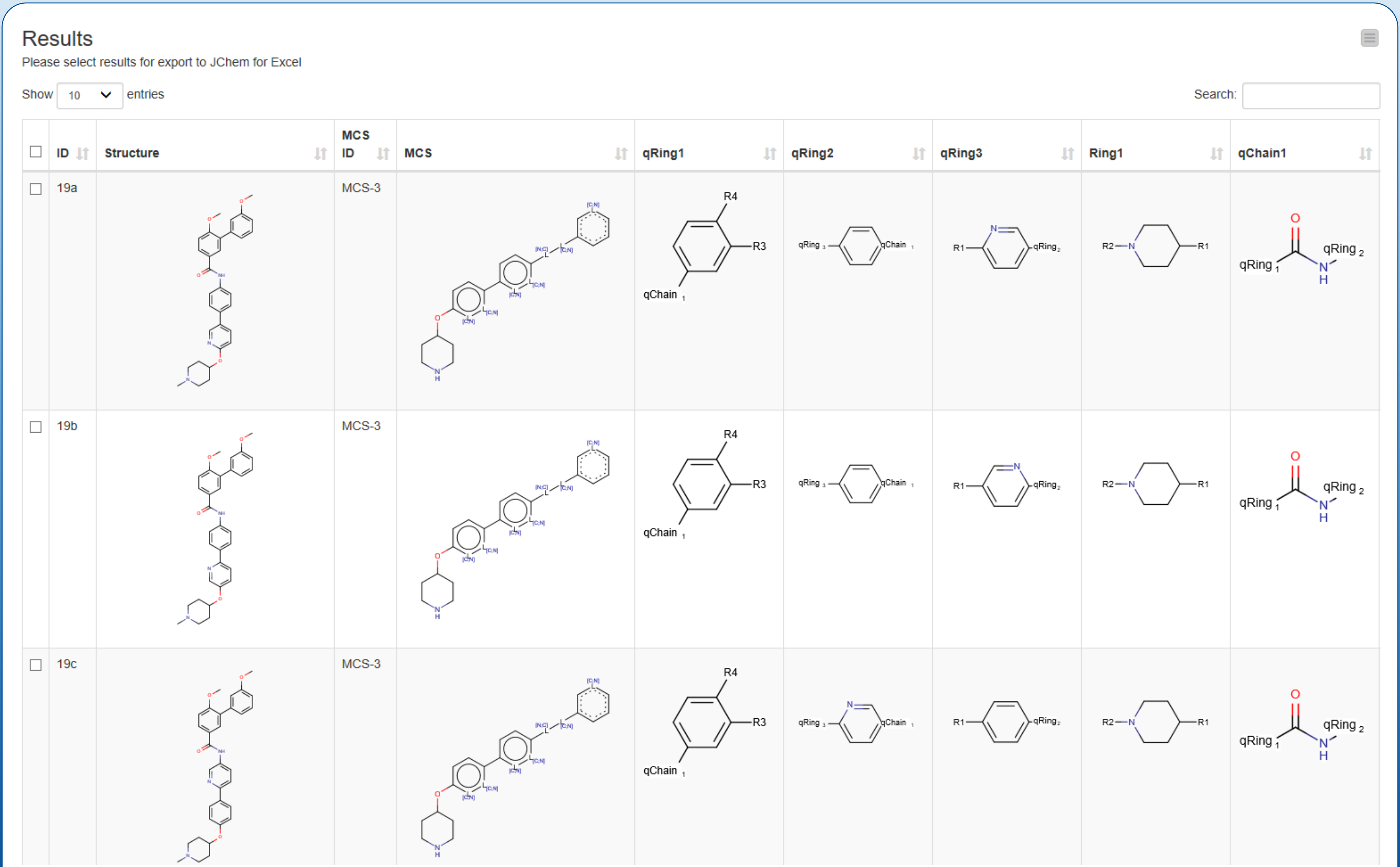
Implementation – Algorithm

- qRings and qChains are either:
  - Expanded to incorporate additional features
  - Fragmented further into new R-groups



Fragment Decomposition

Implementation – Visualisation



Results

Please select results for export to JChem for Excel

Show 10 entries

Search:

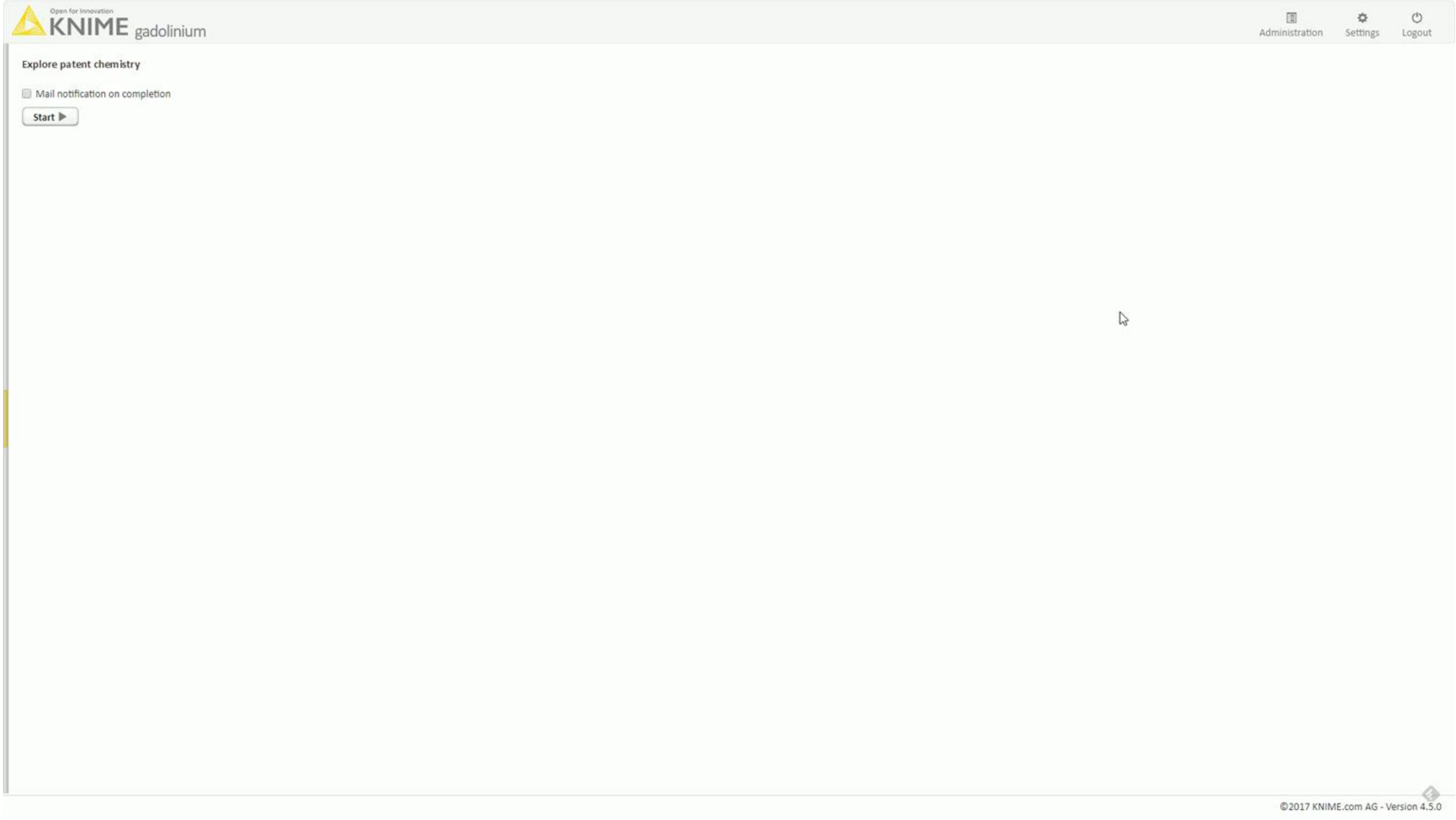
☐	ID	Structure	MCS ID	MCS	qRing1	qRing2	qRing3	Ring1	qChain1
☐	19a		MCS-3						
☐	19b		MCS-3						
☐	19c		MCS-3						

Vernalis
RESEARCH
part of Vernalis plc

Demonstration

THE QUEEN'S AWARDS
FOR ENTERPRISE:
INTERNATIONAL TRADE
2014

Demonstration



Explore patent chemistry

□ Mail notification on completion

Start ▶

©2017 KNIME.com AG - Version 4.5.0

Vernalis
RESEARCH
part of Vernalis plc

Conclusion

THE QUEEN'S AWARDS
FOR ENTERPRISE:
INTERNATIONAL TRADE
2014

Conclusion

Developed tools/algorithms for:

- Extracting structures and data from the patents
- Summarising exemplified structures as fuzzy MCS
- Showing hierarchical relationship between fuzzy MCS
- Performing fragment decomposition driven by fuzzy MCS
- Presenting results to users

Future work:

- More automation of patent document processing
  - Learn lessons from processing more patents
- Multi-parametric SAR/SPR analysis
  - Process and compare multiple tables, e.g. binding and stability data tables
- Integration with other services/tools

Vernalis
RESEARCH
part of Vernalis plc

Thank you!

THE QUEEN'S AWARDS
FOR ENTERPRISE:
INTERNATIONAL TRADE
2014