

The
University
Of
Sheffield.

Representing Clusters of Molecules Using Reduced Graphs

Eleanor Gardiner

University of Sheffield, UK

Outline

- Aims
- Introduction to Reduced Graphs
- Previous work – Virtual screening using RGs
- Cluster representation
- Results
- Conclusions

Aims

- To find a better cluster representative than the fingerprint centroid
- To exploit the generality of Reduced Graphs as molecular descriptors
- To aid in the interpretability of molecular clustering

Why Represent Clusters?

1. To save time when browsing compounds.

A database may have 200 clusters, each with up to 3000 compounds.

2. Related clusters can be identified and merged if necessary.

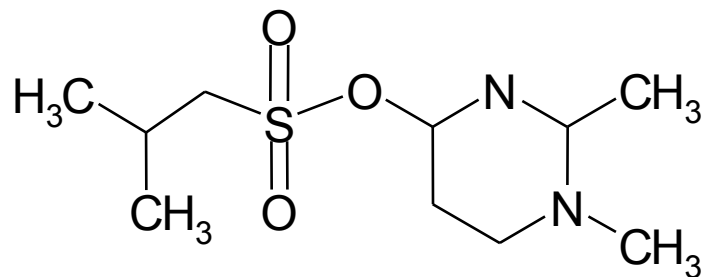
The same core may be present in multiple clusters, especially if initial clustering is done using fingerprints.

Related Work

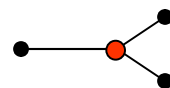
- Database clustering with a combination of fingerprint and maximum common substructure methods¹
 - Uses sphere exclusion, followed by MCS algorithm applied to molecules
 - Clusters are then re-clustered using the cluster representatives

¹Stahl and Mauser(2005) *J Chem Inf Model*

Introduction to Reduced Graphs

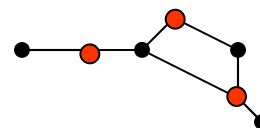


Cyclic/Acyclic Reduction



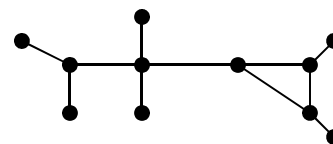
- Ring Node
- Non-ring node

Carbon/Heteroatom Reduction



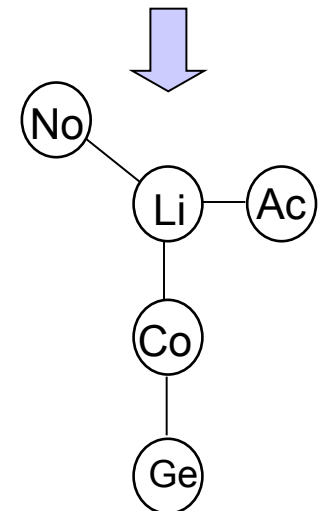
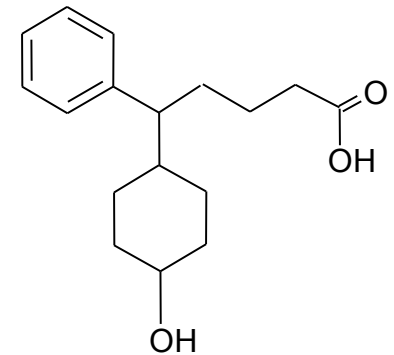
- Heteroatom Node
- Carbon Node

Homeomorphic Reduction



Reduced Graphs

- Pharmacophoric features encoded in 2D
 - Acid; Base; Donor; Acceptor; Aromatic; Alicyclic; Linkers
 - Topology retained
 - No conformational analysis required



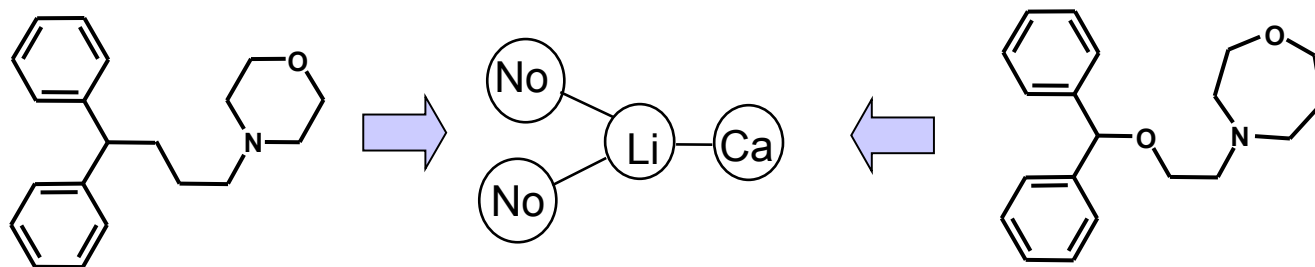
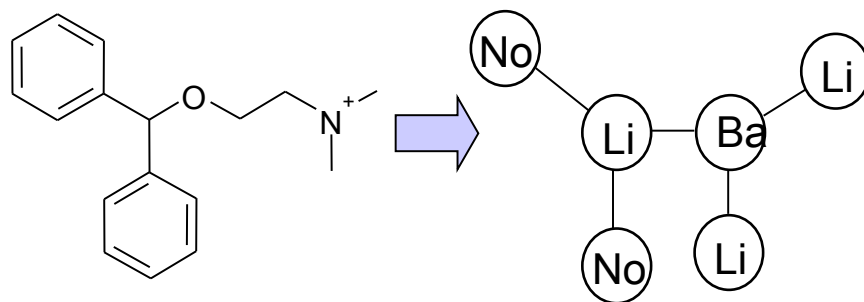
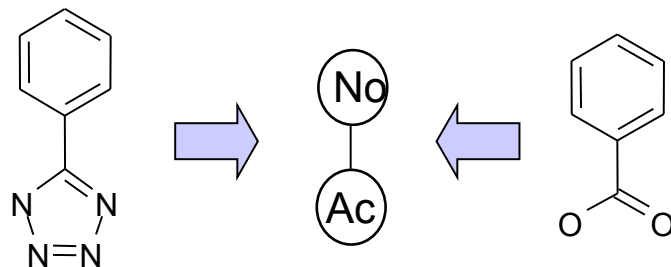
Implementation 1

- Nodes are defined via user-defined SMARTS definitions
- Acid and Base nodes take precedence
- Molecule then partitioned into cyclic and acyclic fragments
- Cycles defined as
 - Aromatic ring; Aromatic ring donor ; Aromatic ring acceptor; Aromatic ring donor and acceptor
 - Aliphatic ring; Aliphatic ring donor ; Aliphatic ring acceptor; Aliphatic ring donor and acceptor
- Acyclic fragments defined as
 - Donor; Acceptor; Donor and acceptor; Linker
- Smiles/SMARTS are parsed using the OEChem toolkit from Openeye

Implementation 2

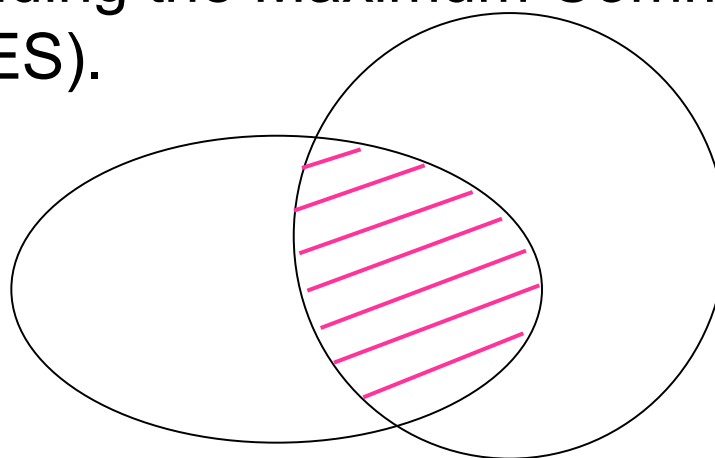
Node description	Node Code
Acyclic inert node.	Li
Acyclic feature node – acceptor only.	Ga
Acyclic feature node – donor only.	Gd
Acyclic feature node – both donor and acceptor.	Ge
Aromatic ring – no hydrogen bonding.	No
Aromatic ring – hydrogen bond acceptor.	Na
Aromatic ring – hydrogen bond donor.	Nd
Aromatic ring – both donor and acceptor.	Ne
Aliphatic ring – hydrogen bond donor.	Cd
Aliphatic ring – hydrogen bond acceptor.	Ca
Aliphatic ring – no hydrogen bonding.	Co
Aliphatic ring – both donor and acceptor.	Ce
Acid feature.	Ac
Base feature.	Ba

Example RGs



Comparing Reduced Graphs

- Compare two RGs by finding the Maximum Common Edge Substructure (MCES).



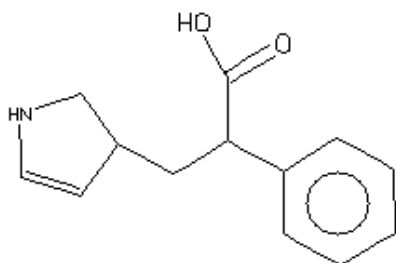
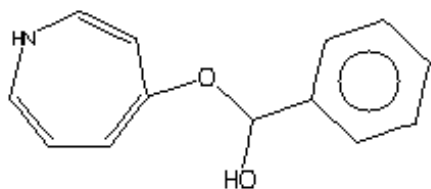
- A measure of the size of the MCES gives a similarity score
- We have used the Rascal^{1,2} algorithm for the MCES

¹Raymond, JW, Willett, P *Journal of Computer-Aided Molecular Design* **2002**, 16, 521-533.

²Raymond, JW, Gardiner, EJ, Willett, P *Computer Journal* **2002**, 45, 631-644.

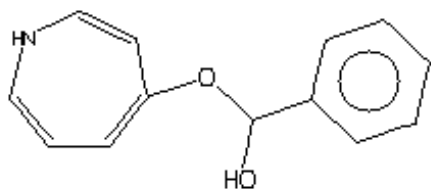
Comparing RGs using MCES and Rascal

Chemical Graph

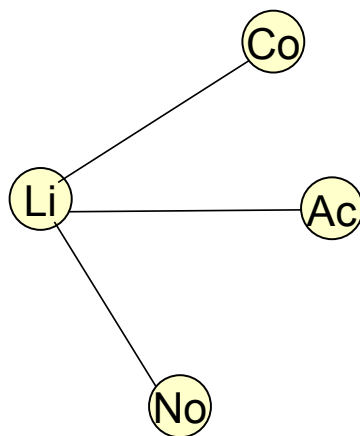
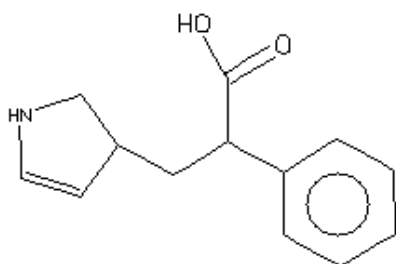
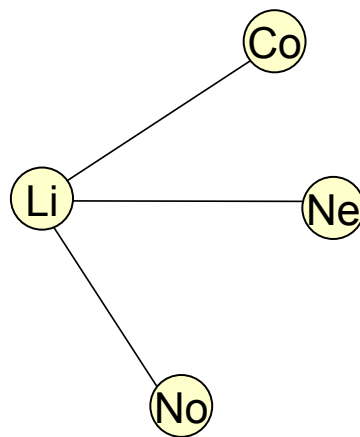


Comparing RGs using MCES and Rascal

Chemical Graph

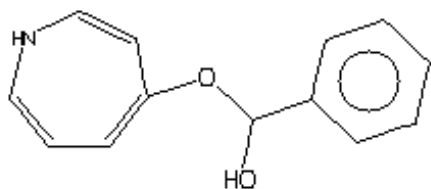


Reduced Graph

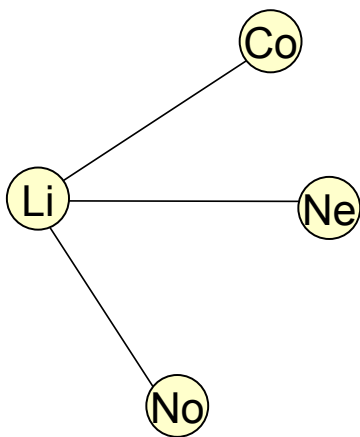


Comparing RGs using MCES and Rascal

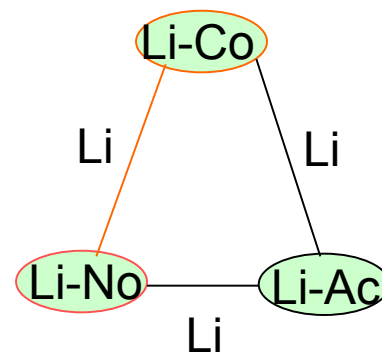
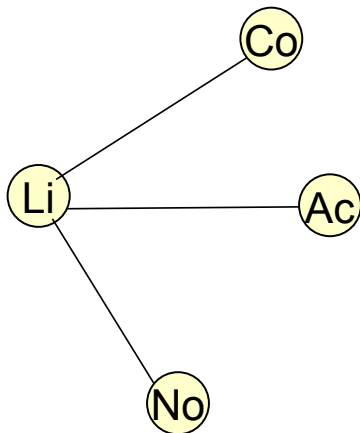
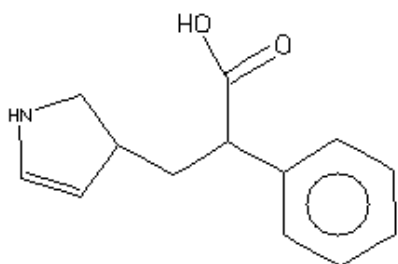
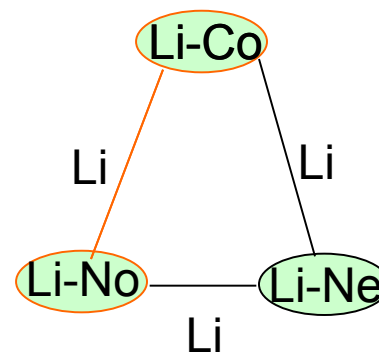
Chemical Graph



Reduced Graph

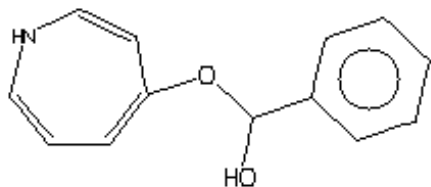


Line Graph

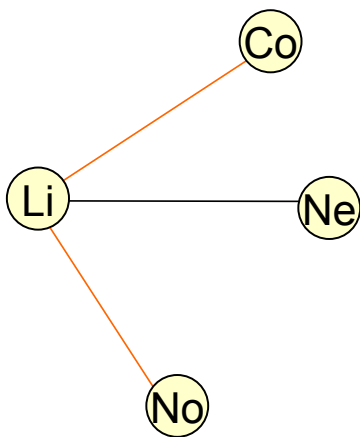


Comparing RGs using MCES and Rascal

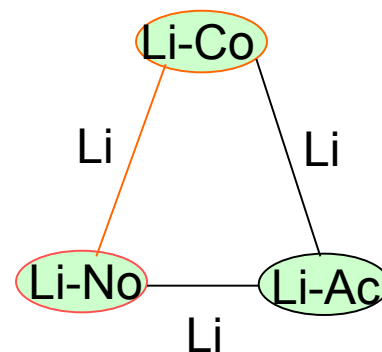
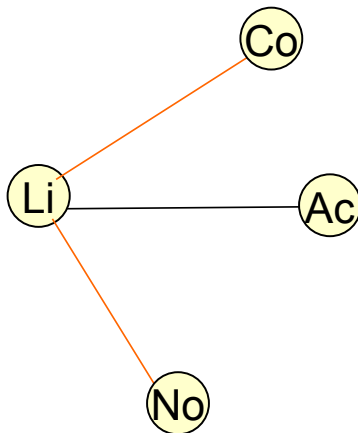
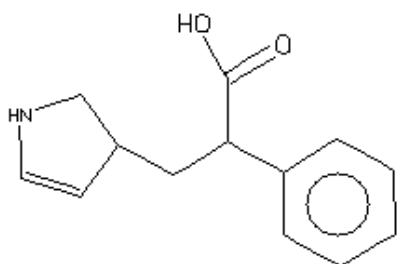
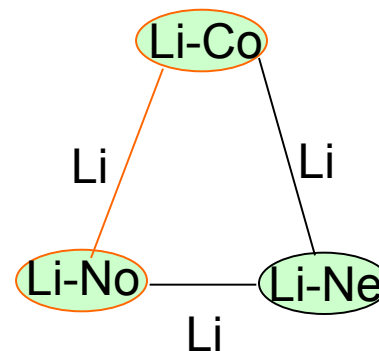
Chemical Graph



Reduced Graph



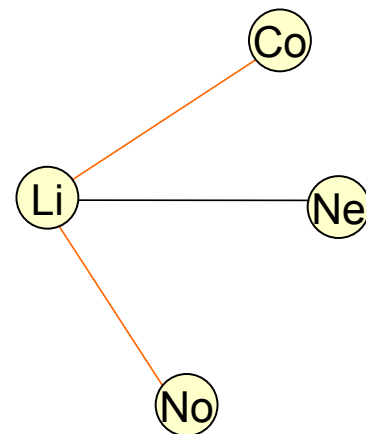
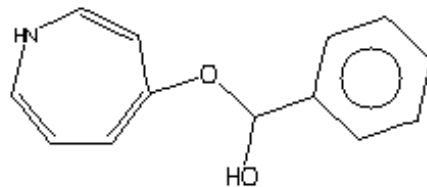
Line Graph



Comparing RGs using MCES and Rascal

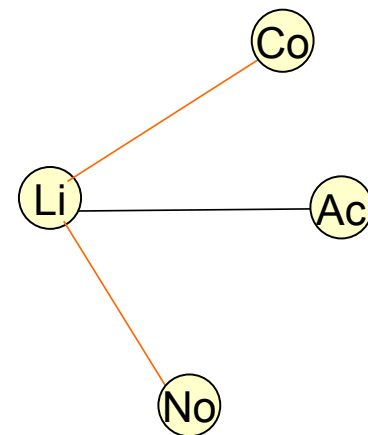
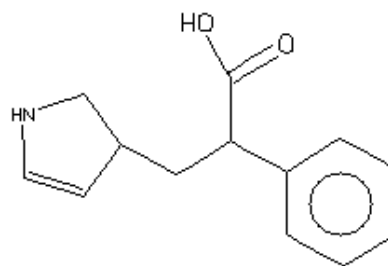
Chemical Graph

Reduced Graph



$$\text{Similarity}_{AB} = \frac{(|V(G_{AB})| + |E(G_{AB})|)^2}{(|V(G_A)| + |E(G_A)|) \cdot (|V(G_B)| + |E(G_B)|)}$$

$$= \frac{(3 + 2)^2}{(4 + 3)(4 + 3)} = 0.51$$



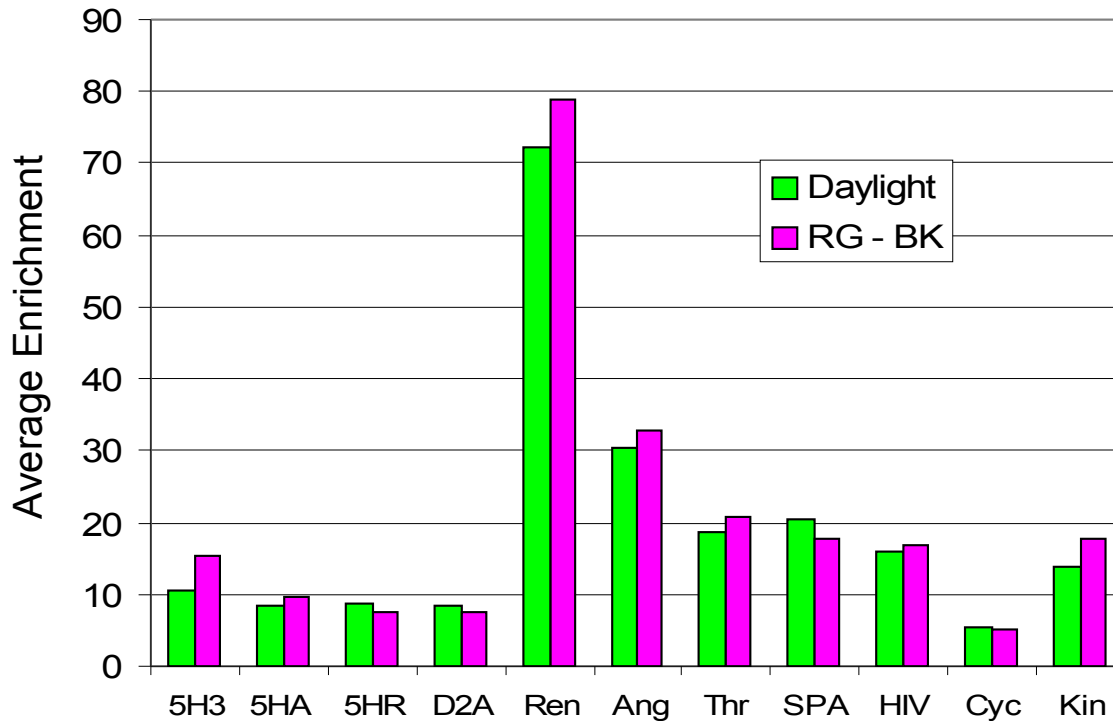
Matching Donors/Acceptor

- Joint donor and acceptor properties (D/A nodes) are allowed to match donors or acceptors, i.e.
 - Ge can match Ga or Gd
 - Ce can match Ca or Cd
 - Ne can match Na or Nd

Previous Virtual Screening Experiments

- MDL Drug Data Report MDDR data put through set of filters
 - Molecular properties (molecular weight, logP, number of rotatable bonds, etc)
 - SMARTS rules obtained from survey of medicinal chemists
 - Remaining (non-ugly) 61902 compounds
- Simulated virtual screening using 11 activity classes
- Compared 4 different RG matching methods

Comparison of RG/Bron-Kerbosch with Daylight Fingerprints



Average Enrichment for 1% recall	
Daylight	19.4
RG BK	20.9
RG FP	19.9
RG Rascal	16.4

Daylight fingerprints and fully connected RG/Bron-Kerbosh are broadly comparable, with the latter slightly better

Why Use Reduced Graphs with Rascal?

- Interpretable – clear connection back to original molecules
- MCES is a molecule and so can be used in a substructure search
- Molecules already clustered – context quite different to virtual screening
- Simplicity

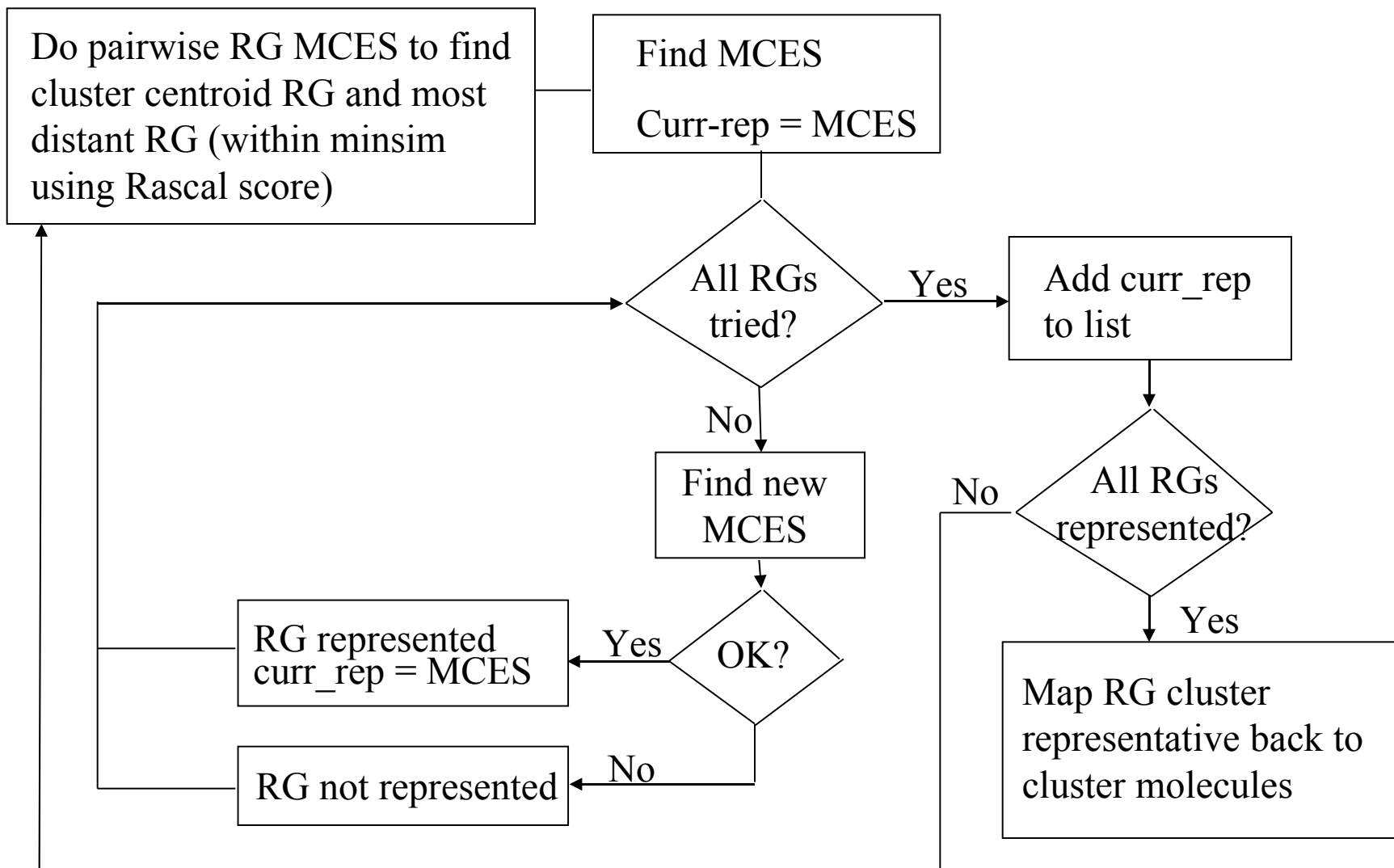
Representing Clusters

- Cluster MDDR using fingerprints, sphere exclusion, Tanimoto score.
- Represent all molecules as RGs
- Process each cluster separately

Initial Clustering Using Sphere Exclusion and Daylight fingerprints

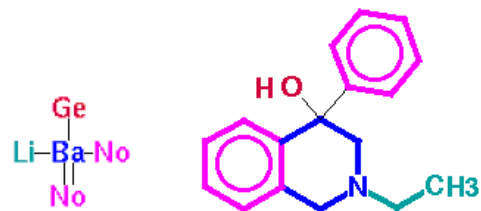
- The cleaned MDDR was clustered using a Tanimoto cutoff of 0.7, giving 14901 clusters, of which ~6000 were singletons (discarded).
- 55315 compounds in non-singleton clusters
- Count the number of unique RGs in each cluster (just regarding them as SMILES strings) – is this all the clustering needed?
 - NO, 38000 unique RGs in ~ 9000 clusters

For Each Cluster

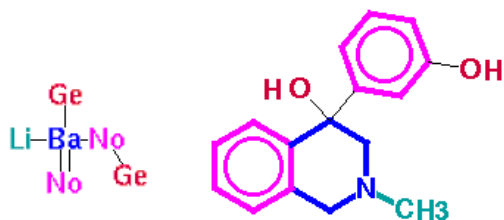


Example - Cluster 904

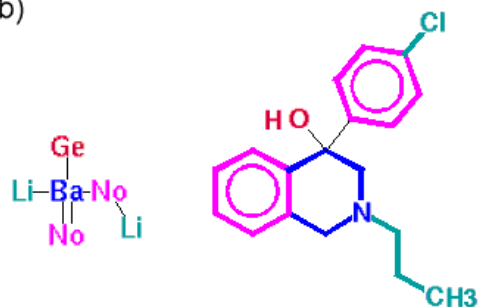
(a)



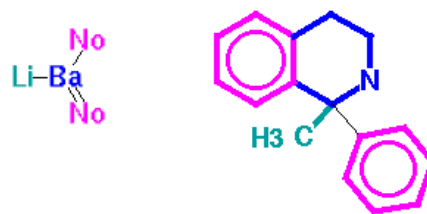
(d)



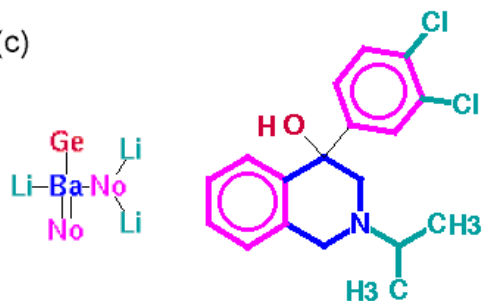
(b)



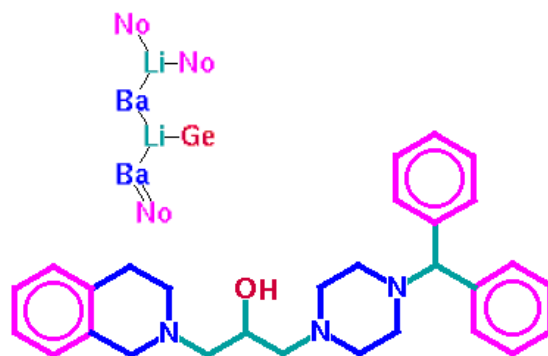
(e)



(c)



(f)



17 molecules,
6 unique RGs

First Find Centroid

17 molecules,

6 unique RGs

Perform pairwise MCES calculations

RGs (a)-(e) each have the same
number (four) RG neighbours

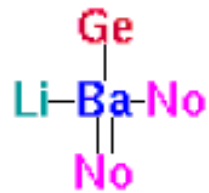
Choose RG (a) as RG centroid

RG (c) is most distant neighbour
within 0.5

Find MCES between RG (a) and
RG(c)

Finding Initial MCES

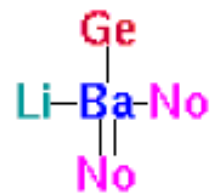
(a)



(c)



Curr_rep



Finding Cluster Representative

Curr_rep



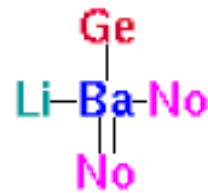
Is a subgraph of

(b)



Finding Cluster Representative

Curr_rep



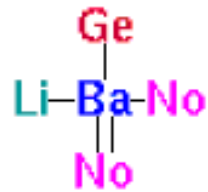
Is a subgraph of

(d)

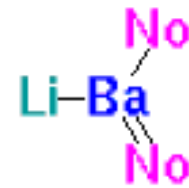


Finding Cluster Representative

Curr_rep

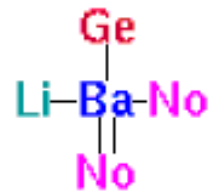


(e)



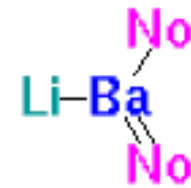
Finding Cluster Representative

Curr_rep



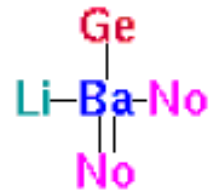
Find MCES

(e)



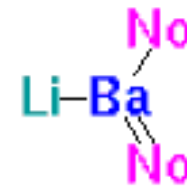
Finding Cluster Representative

Curr_rep

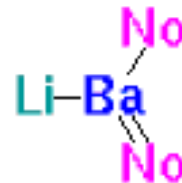


(e)

Find MCES

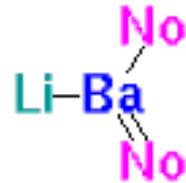


New curr_rep



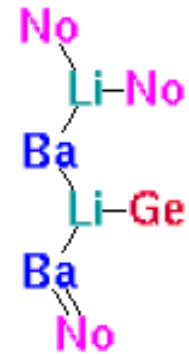
Finding Cluster Representative

Curr_rep



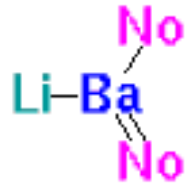
Find MCES

(f)

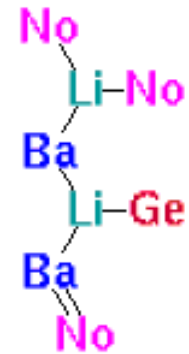


RG Not Represented

Curr_rep



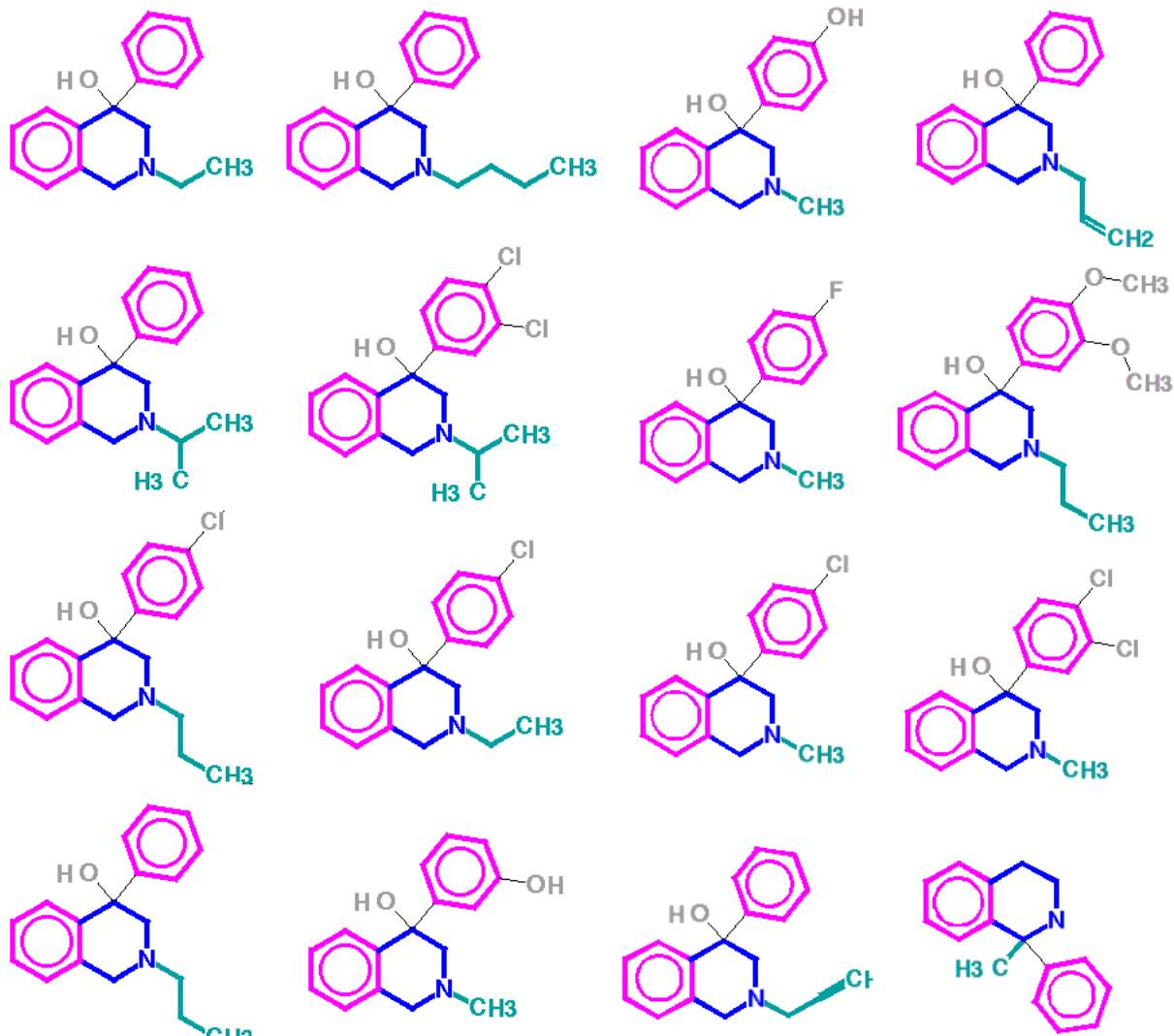
(f)



Similarity too low

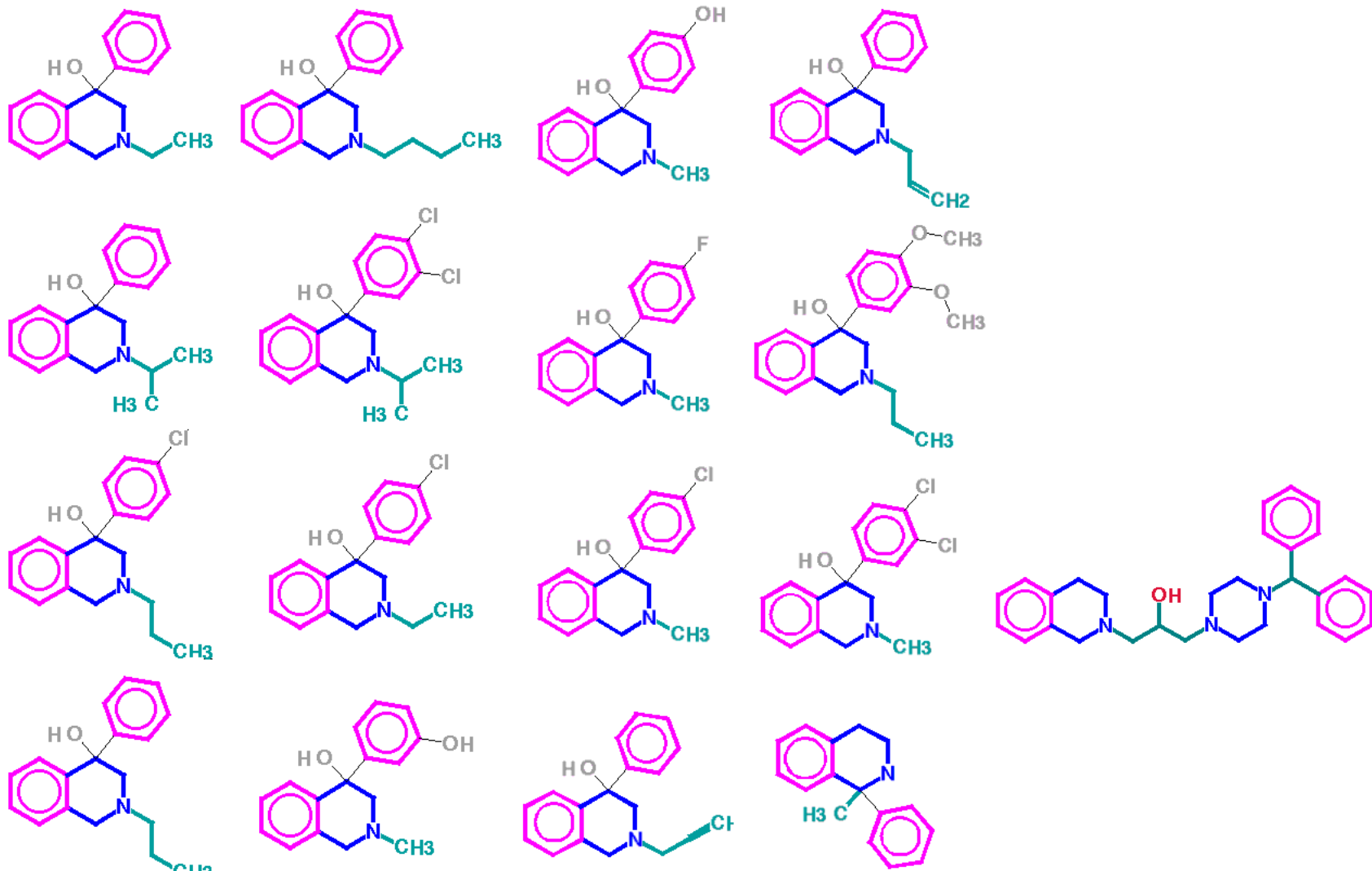
Mapping Cluster Representative to Molecules

No
Li-Ba
No



Mapping Cluster Representative to Molecules

No
Li-Ba
No

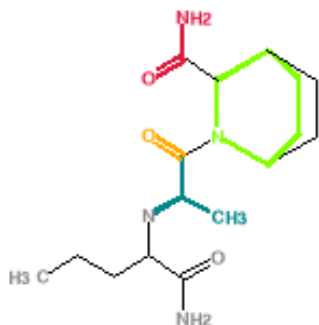


Results Summary

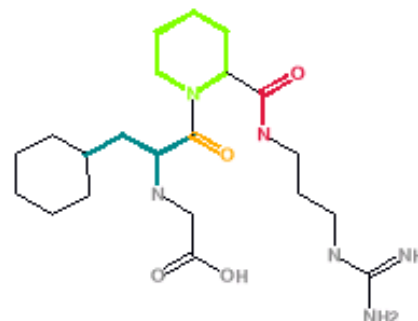
- First 3000 MDDR clusters only
- RASCAL minsim 0.5
- minimum of 3 nodes in cluster rep

Mols represented	33336
Mols per rep	7.65
Reps per cluster	1.58
Clusters with single rep	1465
Unique cluster reps	3945
Total cluster reps	4358
RG nodes per rep	6.55
Clusters not represented	73

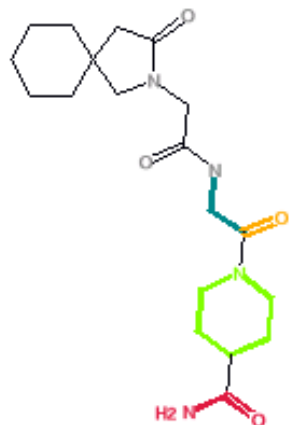
Clusters not Represented



178027



201823

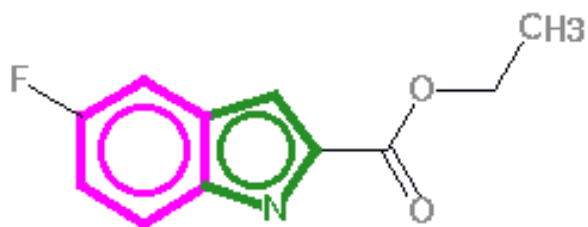


151557

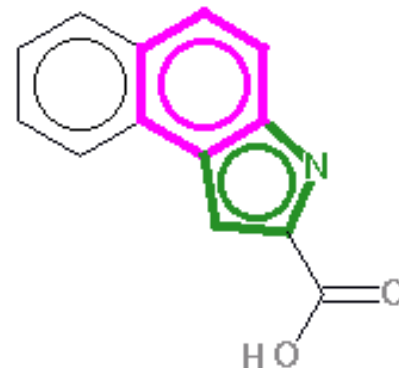


Minimum similarity too high

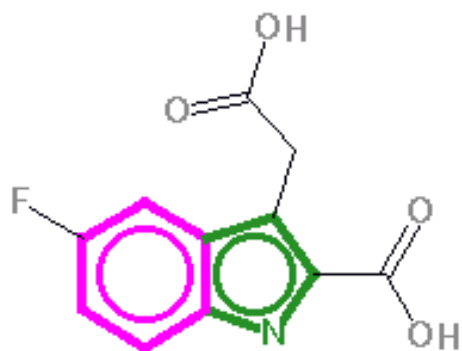
Clusters not Represented



215035



275186

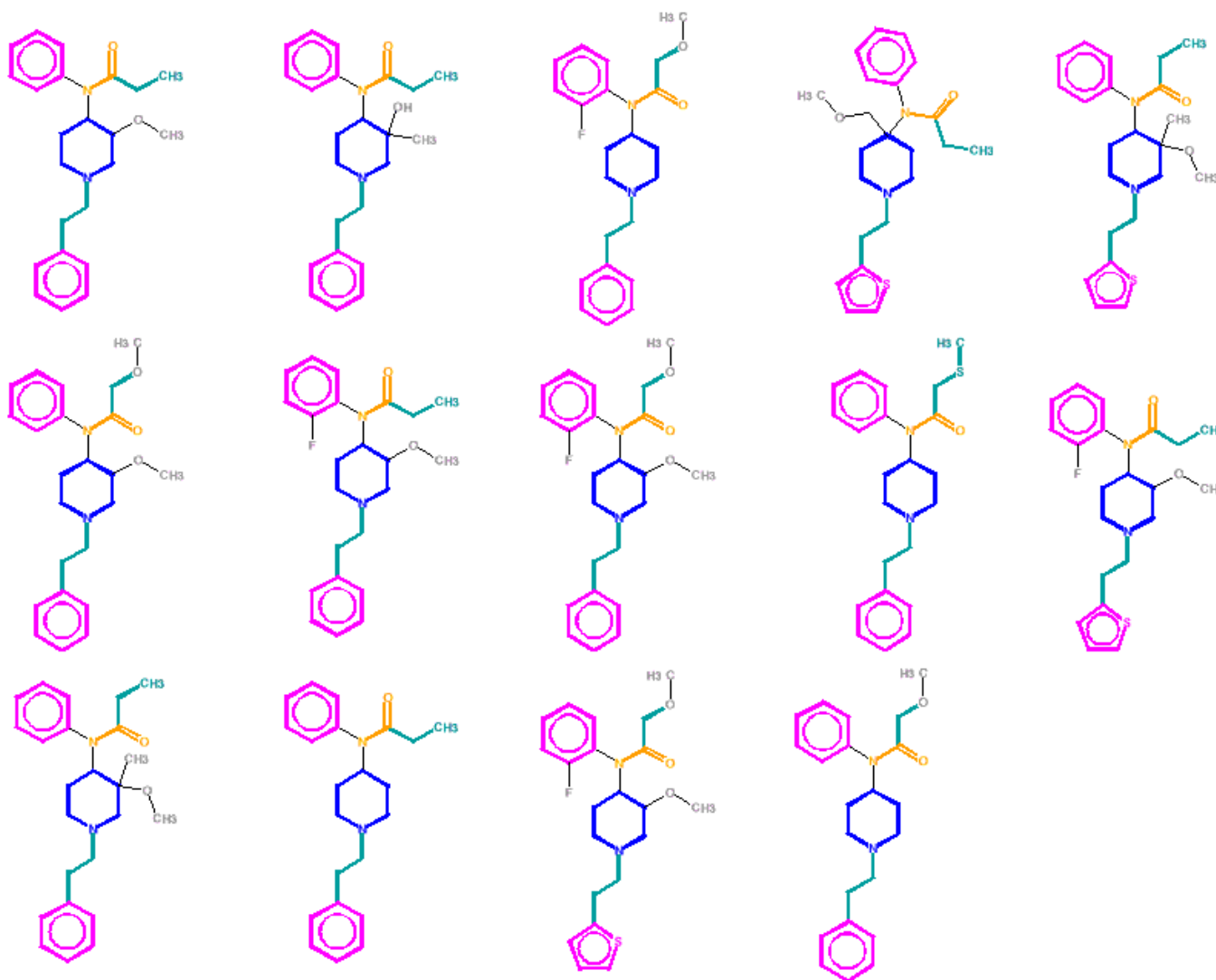
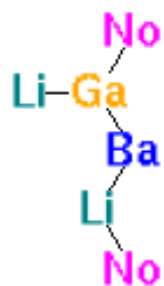


170158

Nd:No

Minimum bonds too small

Cluster 688



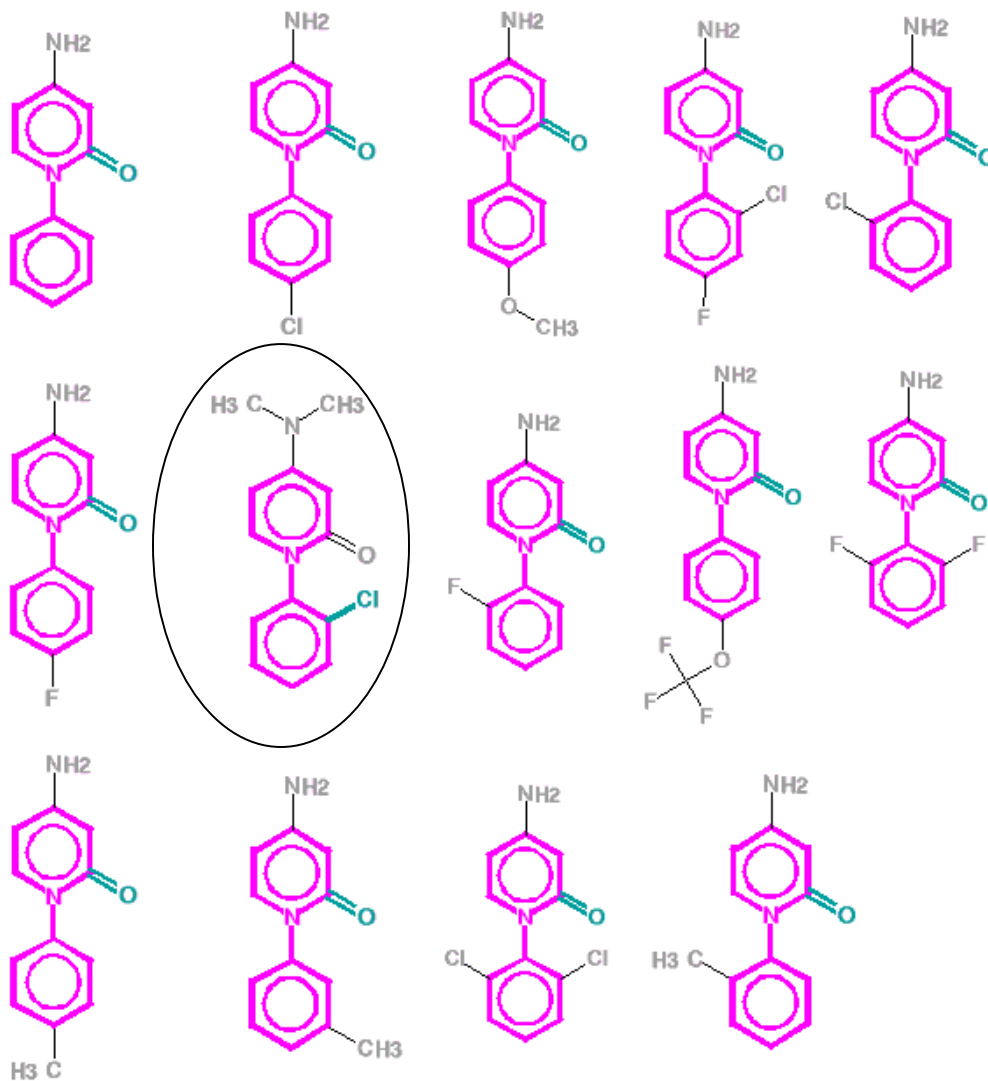
Linker
Acyclic acceptor
Base
Aromatic ring

Caveats

- Of the 3000 MDDR clusters considered, 8 took too long for interactive viewing
- Sometimes the cluster representative is
 - Composed largely of linkers
 - Composed of disconnected fragments
- There may be more than 1 MCES – we choose the first which may not be the ‘best’
- There is not always a unique mapping back to the molecules

Alternative Mapping to Molecules

Li-No
No



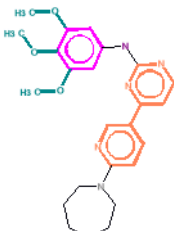
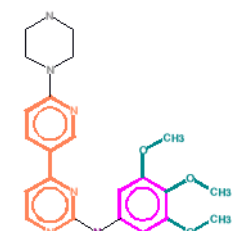
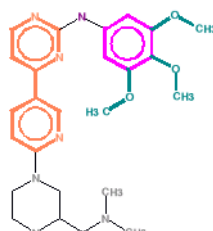
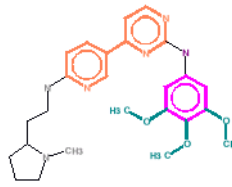
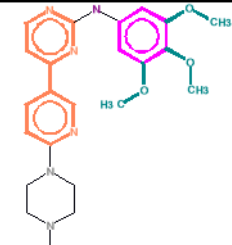
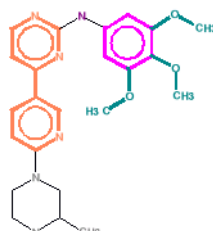
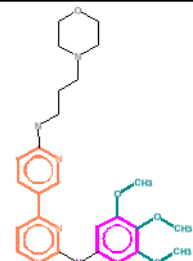
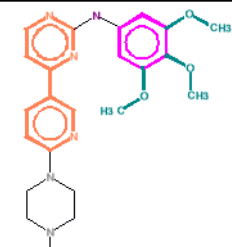
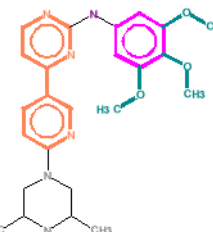
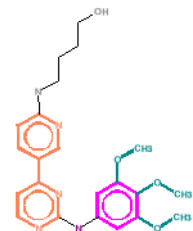
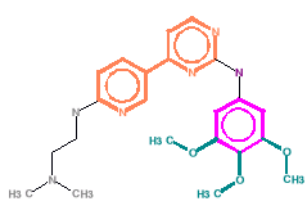
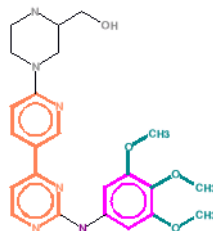
Caveats

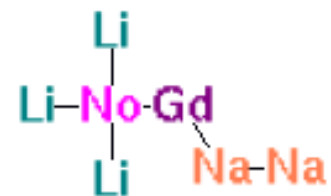
- Of the 3000 MDDR clusters considered, 8 took too long for interactive viewing
- Sometimes the cluster representative is
 - Composed largely of linkers
 - Composed of disconnected fragments
- There may be more than 1 MCES – we choose the first which may not be the ‘best’
- There is not always a unique mapping back to the molecules
- Cluster representation is an order-dependent process

Finding Series

- The initial clustering may group more than one series into a cluster.
- Finding more than one cluster representative can speedily separate them out.
- Example: Cluster 1047 has 16 molecules, 2 representatives

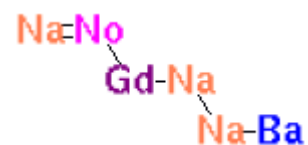
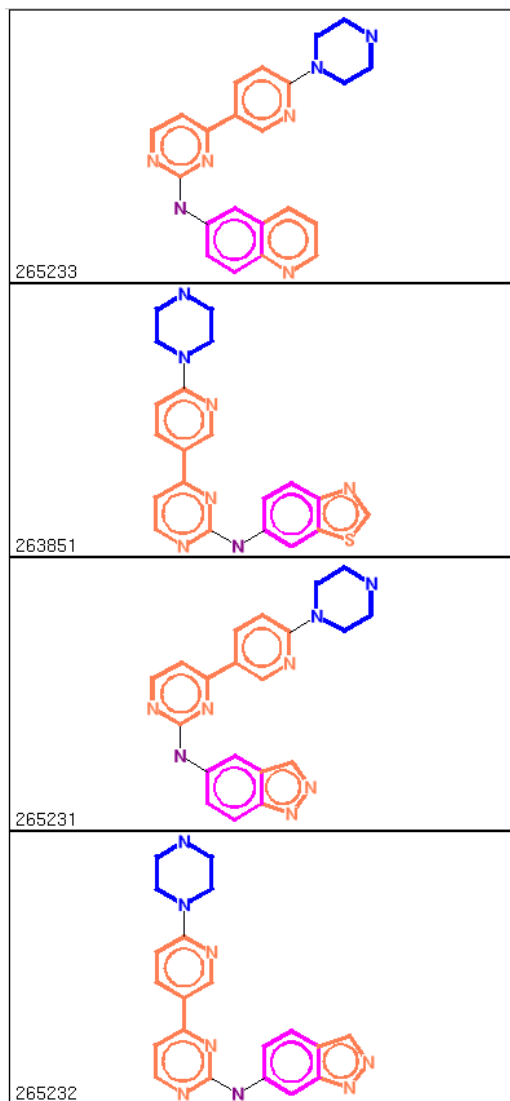
Finding Series

 <p>268136</p>	 <p>265115</p>	 <p>268144</p>
 <p>253058</p>	 <p>268141</p>	 <p>268140</p>
 <p>253057</p>	 <p>268137</p>	 <p>268142</p>
 <p>253054</p>	 <p>253053</p>	 <p>268143</p>



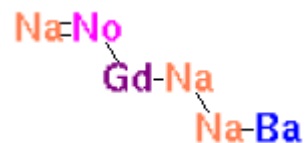
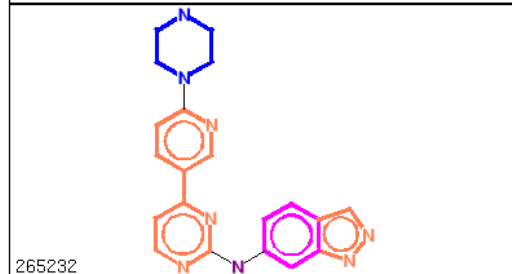
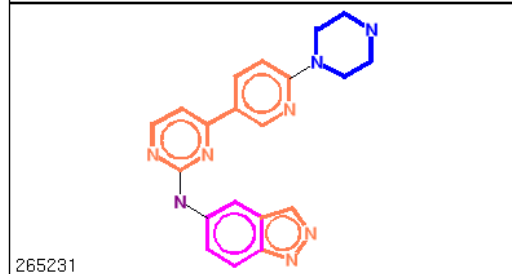
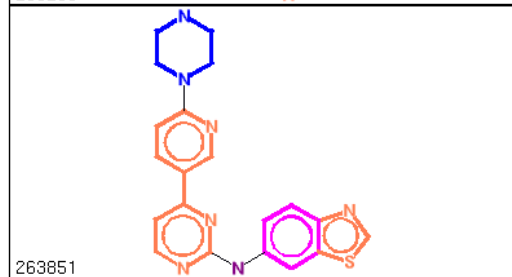
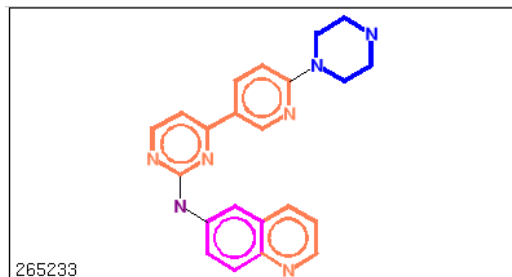
Linker
Aromatic ring
Acyclic donor
Aromatic ring acceptor

Finding Series

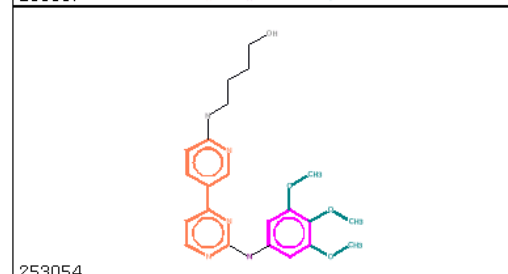
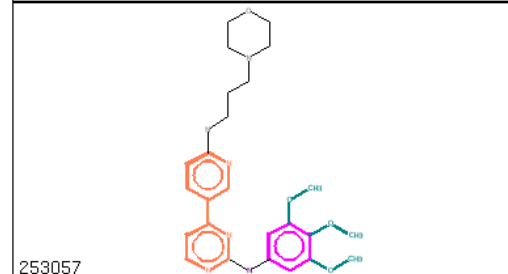
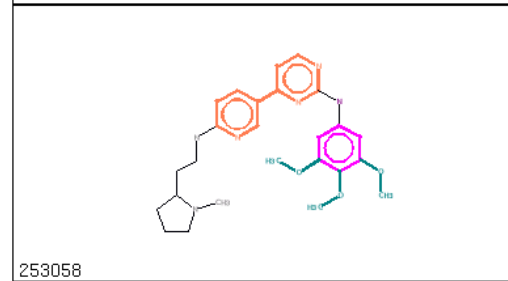
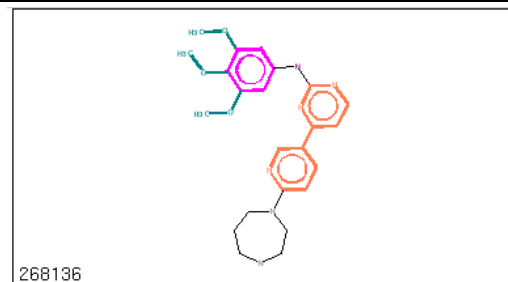


Aromatic ring acceptor
Aromatic ring
Acyclic donor
Base

Finding Series

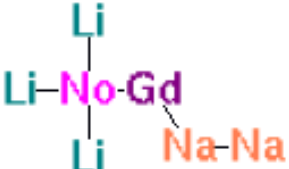


Aromatic ring acceptor
Aromatic ring
Acyclic donor
Base



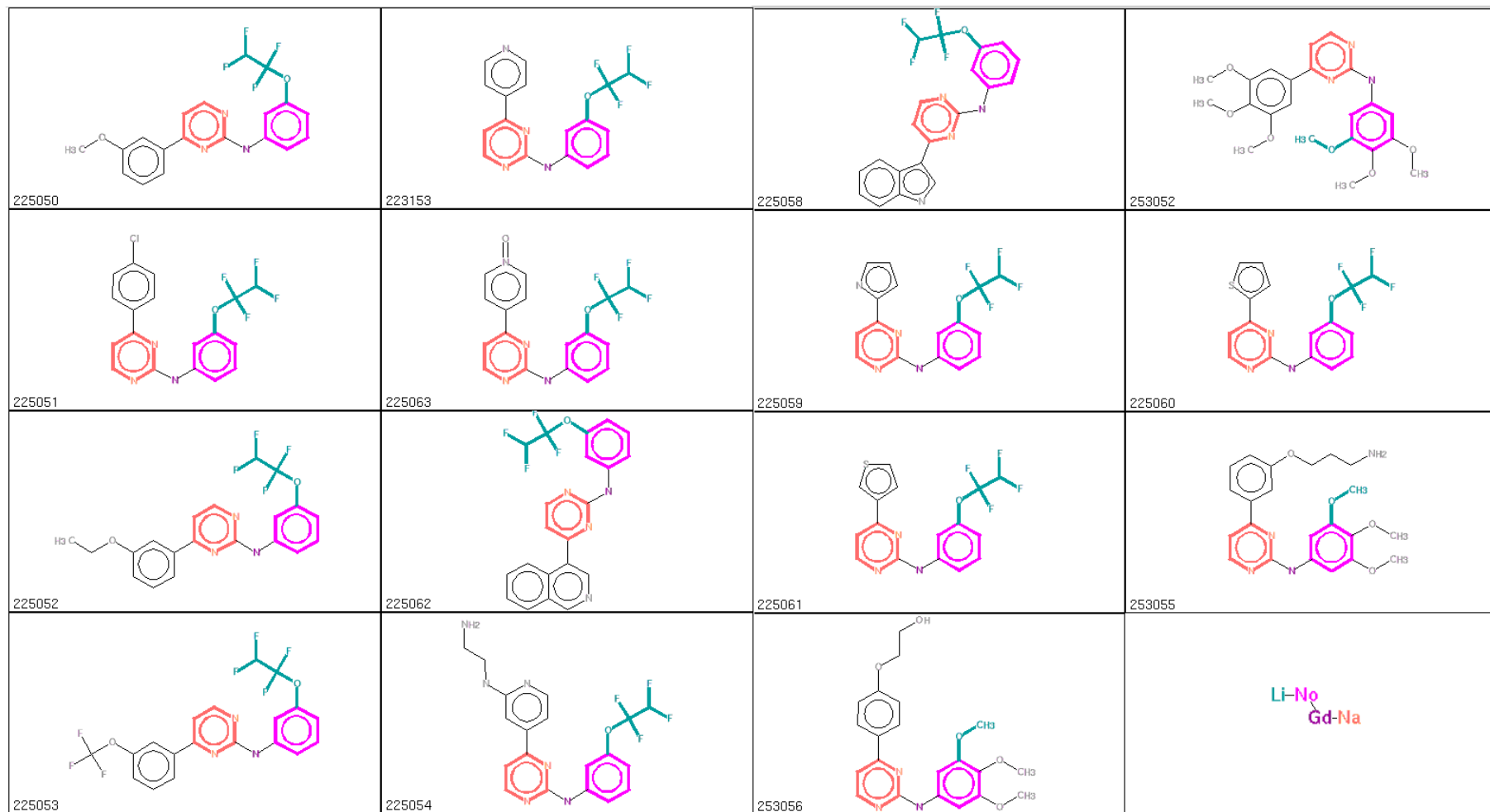
Find Similar Molecules

Use RG cluster representative to search for similar molecules

We used  in RASCAL search of the MDDR for RGs with minimum similarity 0.5.

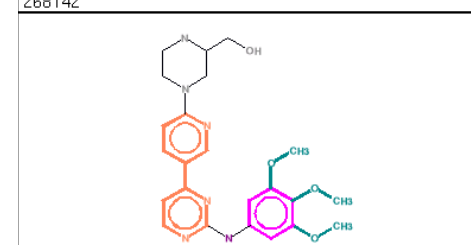
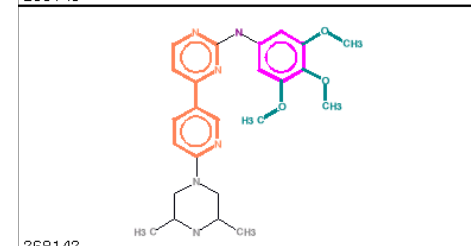
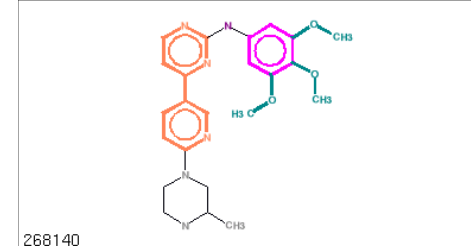
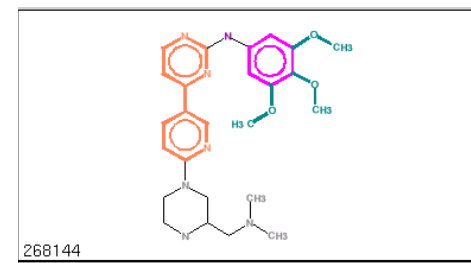
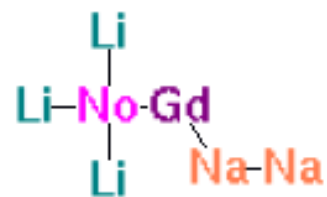
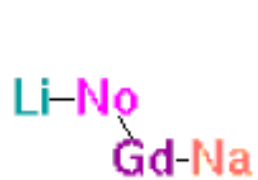
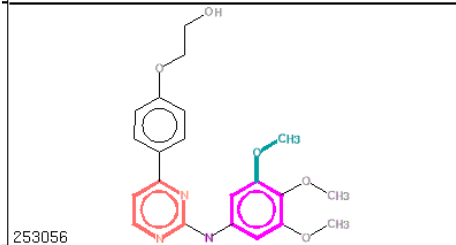
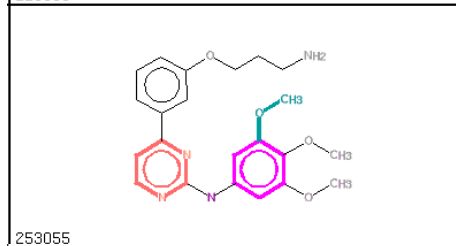
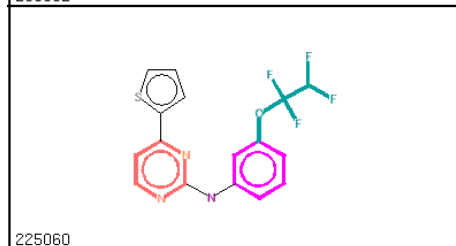
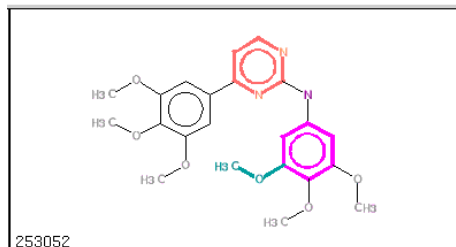
Found 380 molecules including all of cluster 1146 (15 molecules, 11 RGs)

Cluster 1146



1146

1047



Conclusions

- Reduced Graph MCES provides a novel method of cluster representation.
- Fast enough to be used interactively.
- The interpretability of the RG provides a clear link from the cluster representative to the clustered molecules.
- RG cluster representatives can clearly indicate the fixed and variable parts of a series of compounds.
- RG cluster representation can divide clusters into two or more series.
- RG cluster representatives can be used to find similar molecules and aid in merging clusters

Acknowledgements

- **Sheffield University**
Val Gillet
Peter Willett
- **AstraZeneca**
David Buttar
Dave Cosgrove
Paula Kitts
- Funding, software and data support: AstraZeneca, Daylight CIS, OpenEye, MDL Information Systems, Royal Society and Wolfson Foundation

Selected Sheffield Reduced Graph Publications

- Holliday, JD et al. (1994) Evaluation of the screening stages of the Sheffield research-project on computer-storage and retrieval of generic chemical structures in patents. *JCICS*, **34**, 39-46.
- Gillet, VJ et al. (2003) Similarity searching using reduced graphs. *JCICS*, **43**, 338-345.
- Barker, EJ et al. (2003) Further development of reduced graphs for identifying bioactive compounds. *JCICS*, **43**, 346-356.
- Barker, EJ et al. (2006) Scaffold hopping using clique detection applied to reduced graphs. *JCIM*, **46**, 503-511.
- Gardiner, EJ et al. (2007) Representing clusters using a maximum common edge substructure algorithm applied to reduced graphs and molecular graphs. *JCIM*, **47**, 354-366.