
Mining for context-sensitive matched molecular pairs and bioisosteric replacements in large chemical databases

George Papadatos

Postdoctoral Research Fellow

Eli Lilly

g.papadatos@lilly.com

Outline

■ Theory

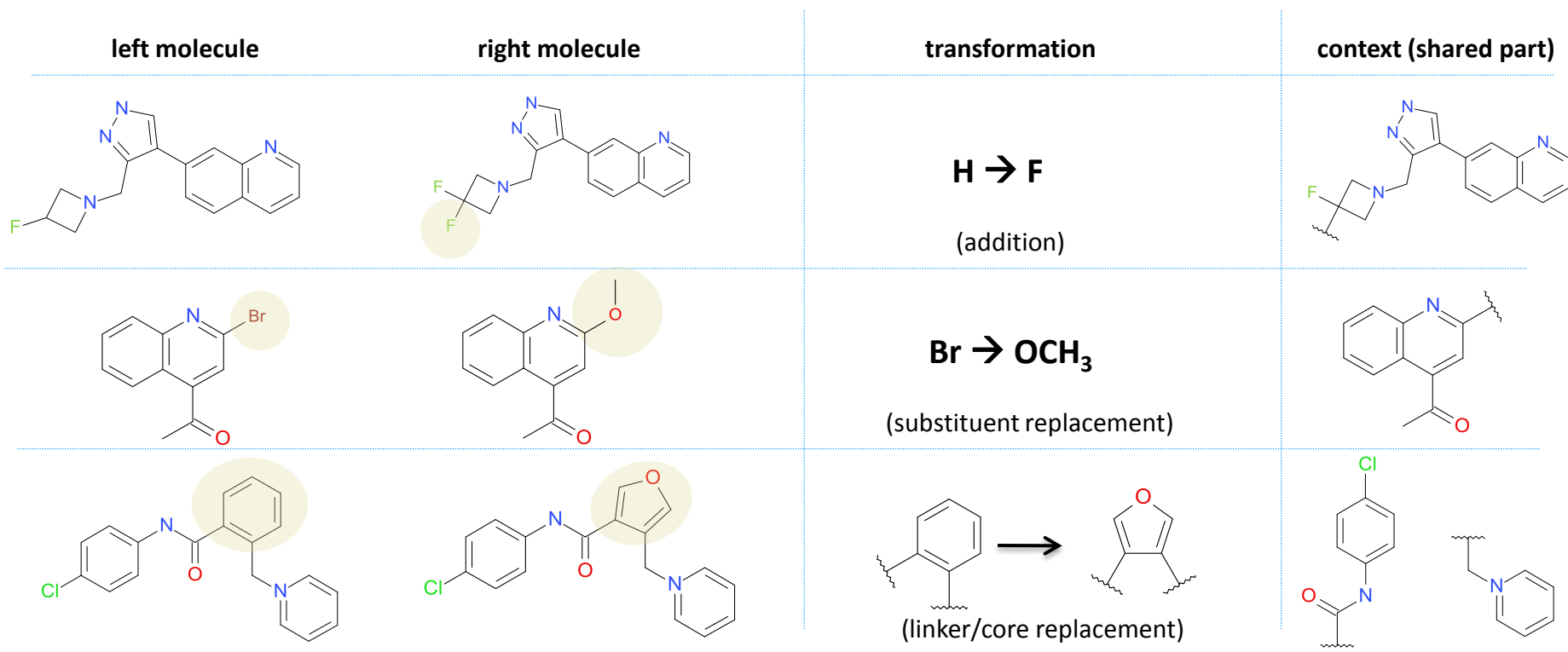
- What is a matched molecular pair (MMP)?
- What is a MMP analysis (MMPA)?
- How do we automatically detect MMPs?
- Why the fuss?

■ Practice

- Investigation of the role of the context (GSK)
- Mining for interesting transformations (Lilly)
- MMP IT integration in Lilly UK
 - Case scenarios: MMP generation and exploitation

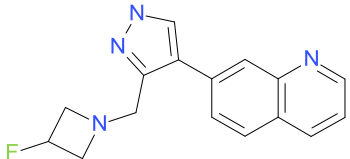
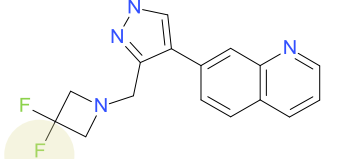
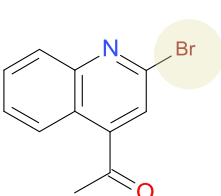
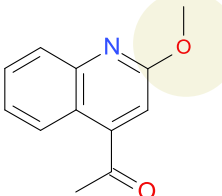
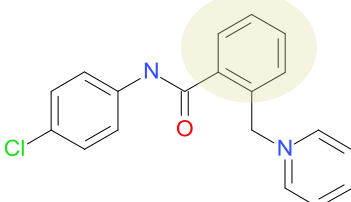
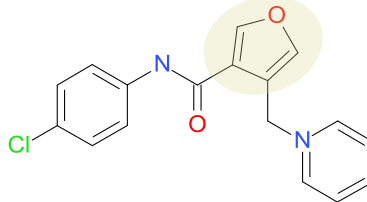
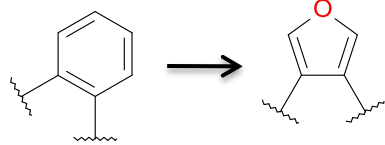
What is a matched molecular pair?

- Two molecules which differ from each other by a specific and small change, while their shared structural part is identical

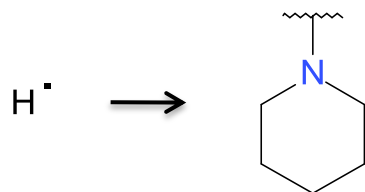


What is MMP analysis (MMPA)?

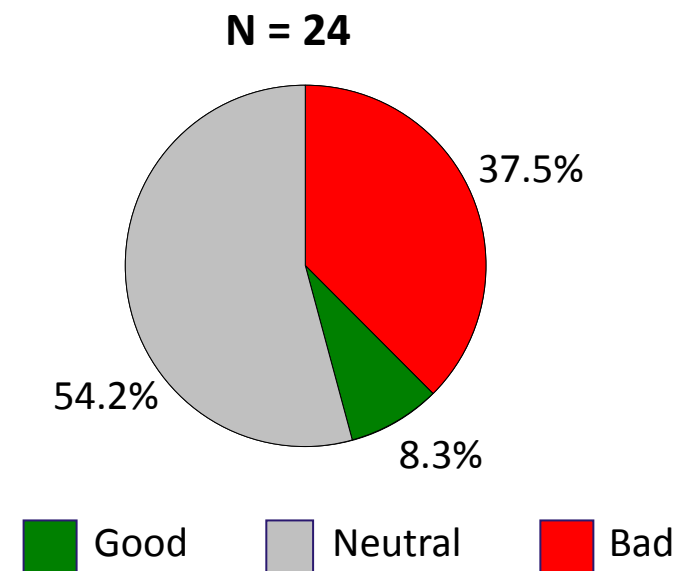
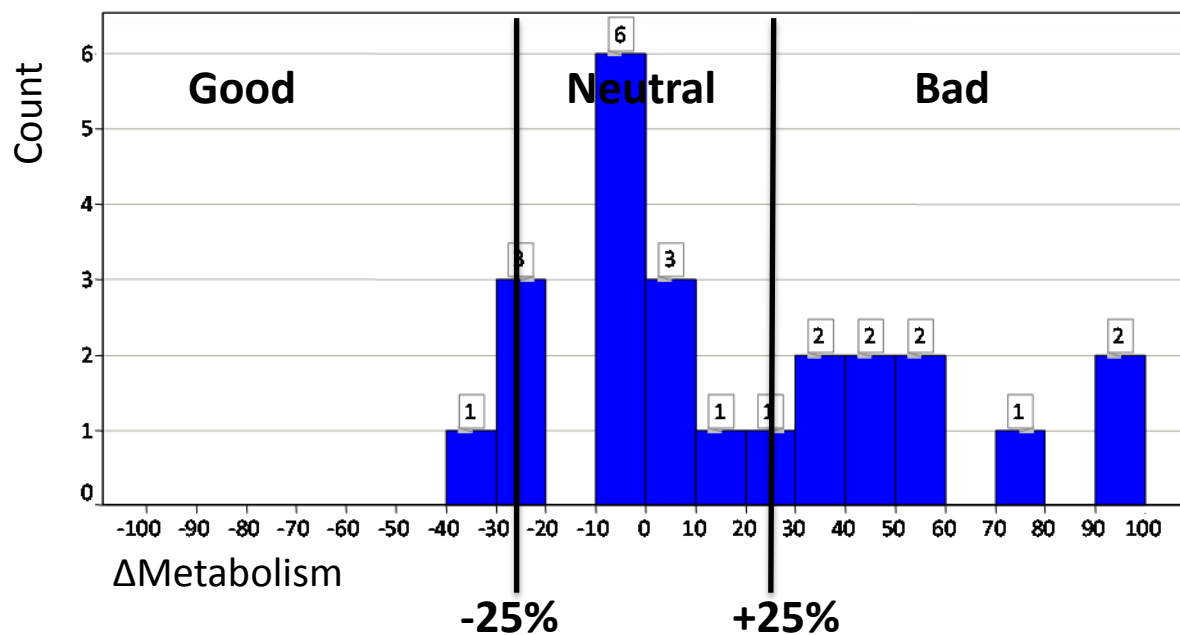
- The mining and statistical analysis of transformations and their impact on properties of interest (e.g. met. stability)

left molecule	right molecule	transformation	Δ Stability (% metabolised)
		$H \rightarrow F$	-30.5
		$Br \rightarrow OCH_3$	+40.7
			+60.3

MMP analysis in action



N = 24 observations (MMPs)

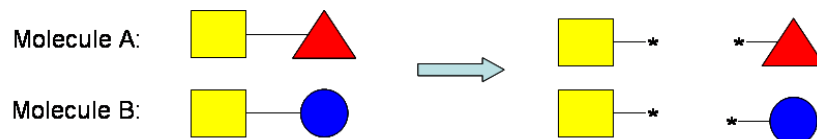


How do we mine for MMPs?*








(*in an automated and unsupervised way)

- It used to be a slow and computationally expensive process...
 - Pair-wise maximum common substructure extraction – $O(N^2)$
- Recently** a much more efficient algorithm was published

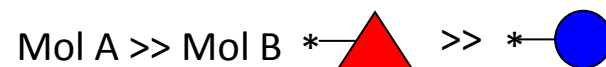
1) Cleave all acyclic single bonds, one by one:



2) Index all the fragments (cf. book index):

Key	Value
 *	 * Mol A  * Mol B
* 	 * Mol A
* 	 * Mol A

3) Enumerate the values for each key:



Hussain and Rea (2010). *J. Chem. Inf. and Model.*, **50** (3), 339-348.

Wagener and Lommerse (2006). *J. Chem. Inf. and Model.*, **46 (2), 677-685.

Why the fuss?

- Intuitive, straightforward, interpretable, inverse-QSAR approach 😊
 - No models, no obscure fingerprints or similarity assessments
- MMPA can answer questions such as:
 - Which are the most frequent transformations? What is the average impact of a transformation on a property? Are there consistently 'good' or 'bad' transformations? What are the novel *bioisosteric* replacements found in kinase inhibitors / in the public domain / patents?
- MMPA can retrieve historical precedents:
 - Get me all the MMPs and transformations that increased solubility **and** clogP
 - Get me all the transformations that improved activity X by > 2 log units
 - I want to replace an ethoxy group and improve met. stability for my project. What has been done in the past (company-wide or in the public domain)? Is there any precedent of a successful replacement for ethoxy?

Papadatos et al. (2010). *J. Chem Inf. and Model.*, **50** (10), 1872-1886.

Gleeson et al. (2009). *Bioorg. & Med. Chem.*, **17** (16), 5906-5919.

Hajduk and Sauer (2008). *J. Med. Chem.*, **51** (3), 553-564.

Warner et al. (2010). *J. Chem Inf. and Model.*, **50** (8), 1350-1357.

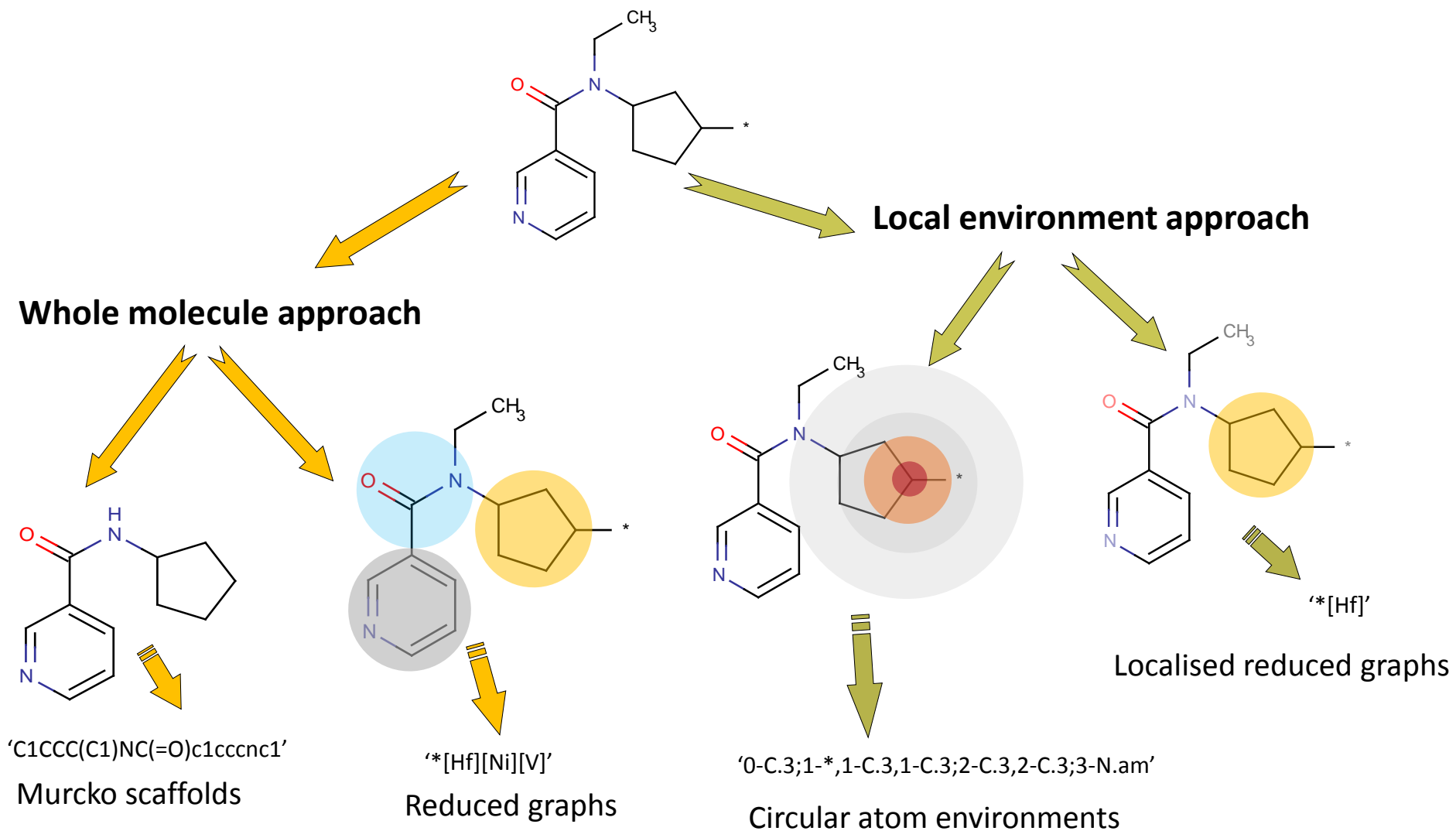
Leach et al. (2006). *J. Med. Chem.*, **49** (23), 6672-6682.

Investigating the role of the context

- Large files of compounds tested in reliable, consistent assays
 - GSK hERG inhibition, aqueous solubility, logD (76K-180K data points)
- No predefined list of transformations
 - Terminal substituent replacements
- Robust statistics based on large no. of examples
- Context-sensitive MMP analysis
 - No assumptions for global effect of a transformation
 - Is the effect the same for local regions of chemical space?
 - Investigation for several descriptors which represent either the whole context or the local environment where the change took place
 - Identify discrepancies between global and local ΔP distributions

Papadatos et al. (2010). *J. Chem Inf. and Model.*, 50 (10), 1872-1886.

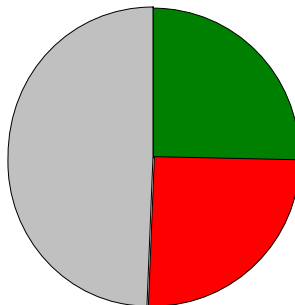
Context representation



H>>OCH3 - hERG

Global Δ hERG distribution

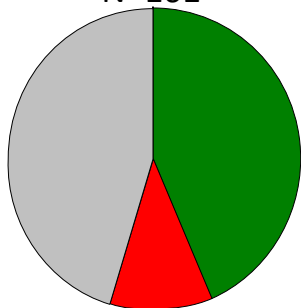
N=2484



vs.

Local Δ hERG distributions – localised reduced graphs

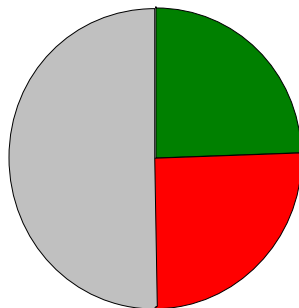
N=161



*[A]: aliphatic linker

p = 3.9E-8

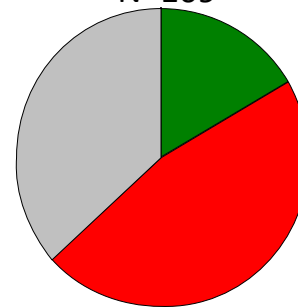
N=2151



*[B]: featureless arom. ring

p = 0.58

N=109



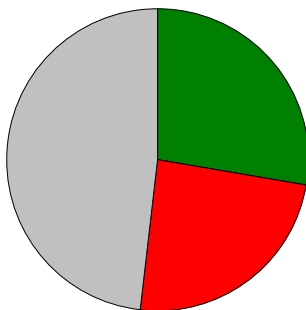
*[C]: H-bond acceptor arom. ring

p = 1.3E-6

Cyclohexyl >> Phenyl - hERG

Global Δ hERG distribution

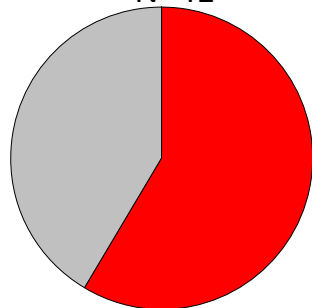
N=303



vs.

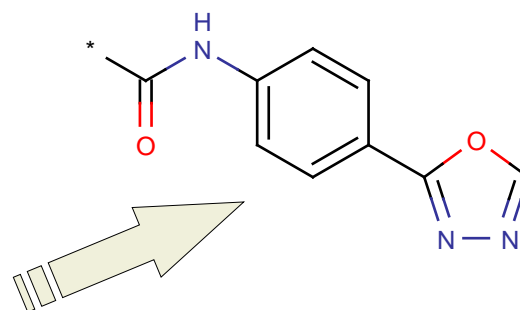
Local Δ hERG distribution – Murcko scaffolds

N=41



*C(=O)Nc1ccc(cc1)c2nnco2

p = 1.8E-8



Summary of results

Cases where $p < 0.01$			
Descriptor	hERG	Solubility	Lipophilicity
Murcko frameworks	159	243	1242
Reduced graphs	165	320	1329
Atom environments	229	274	1719
Localised RG node	32	91	439

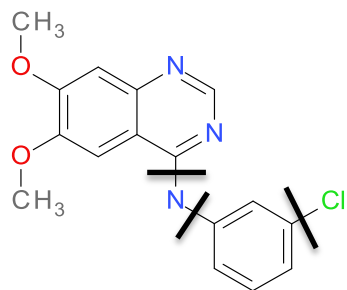
Papadatos et al. (2010). *J. Chem Inf. and Model.*, **50** (10), 1872-1886.

Mining for interesting transformations

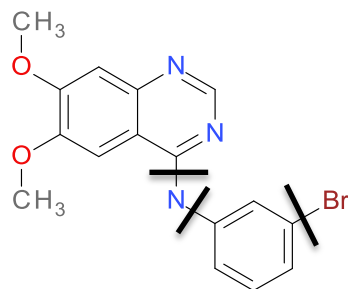
1. Begin with a large dataset with property values
 - 43.3K data points, Lilly human liver microsomal stability assay
2. Find all matched pairs
 - Terminal substituent *and* linker/core replacements
3. Extract contexts and corresponding transforms
4. Canonicalise direction of transformations (i.e. $A \rightarrow B$ equiv. $B \rightarrow A$)
5. Group transformations; calculate and bin ΔP s in 3 bins
6. Identify bioisosteric replacements
 - Majority of Δ Metabolism values should be between -25% and 25%
 - Number of distinct Murcko scaffolds > 10
7. Analyse trends in the effect of each transformation

Context revisited...

left molecule



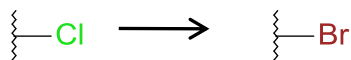
right molecule



Heavy atoms involved

Δ Metabolism

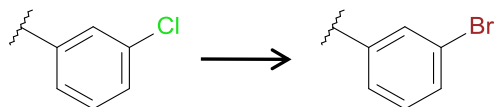
Level 1



2

28

Level 2



14

28

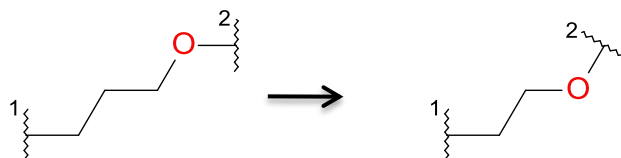
Level 3



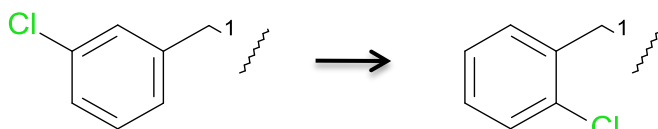
16

28

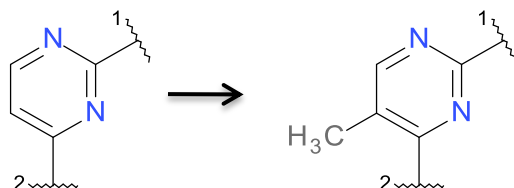
Boring bioisosteric transformations



33 examples found
Mean ΔP : **-1.44%**
Neutral Count Ratio: **100%**



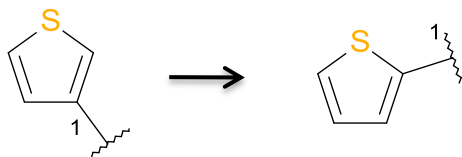
21 examples found
Mean ΔP : **5.51%**
Neutral Count Ratio: **100%**



25 examples found
Mean ΔP : **4.2%**
Neutral Count Ratio: **100%**

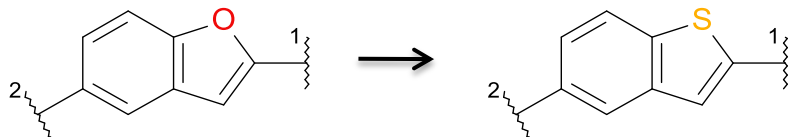


31 examples found
Mean ΔP : **2.26%**
Neutral Count Ratio: **100%**

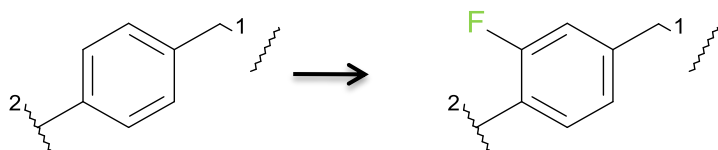


31 examples found
Mean ΔP : **0.87%**
Neutral Count Ratio: **96.7%**

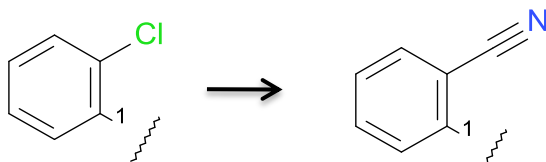
More exciting bioisosteric replacements



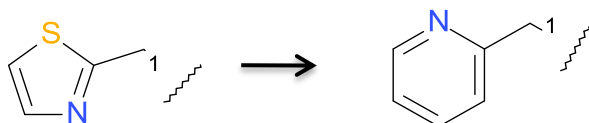
28 examples found | Mean Δ P: **1.9%**
Neutral Count Ratio: **96.3%**



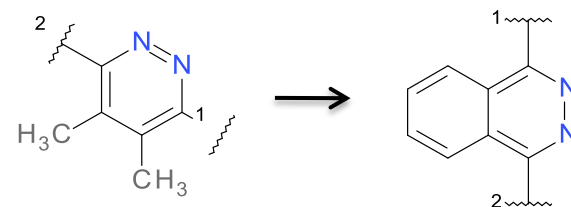
38 examples found | Mean Δ P: **3.41%**
Neutral Count Ratio: **97.3%**



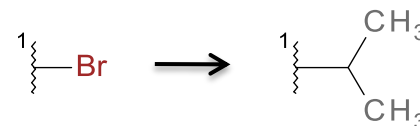
28 examples found | Mean Δ P: **-4.1%**
Neutral Count Ratio: **96.4%**



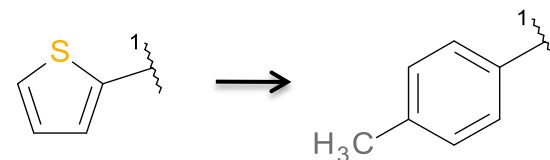
27 examples found | Mean Δ P: **-0.85%**
Neutral Count Ratio: **96.15%**



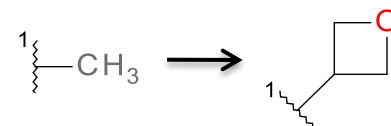
45 examples found | Mean Δ P: **5.30%**
Neutral Count Ratio: **95.5%**



19 examples found | Mean Δ P: **-2.25%**
Neutral Count Ratio: **100%**

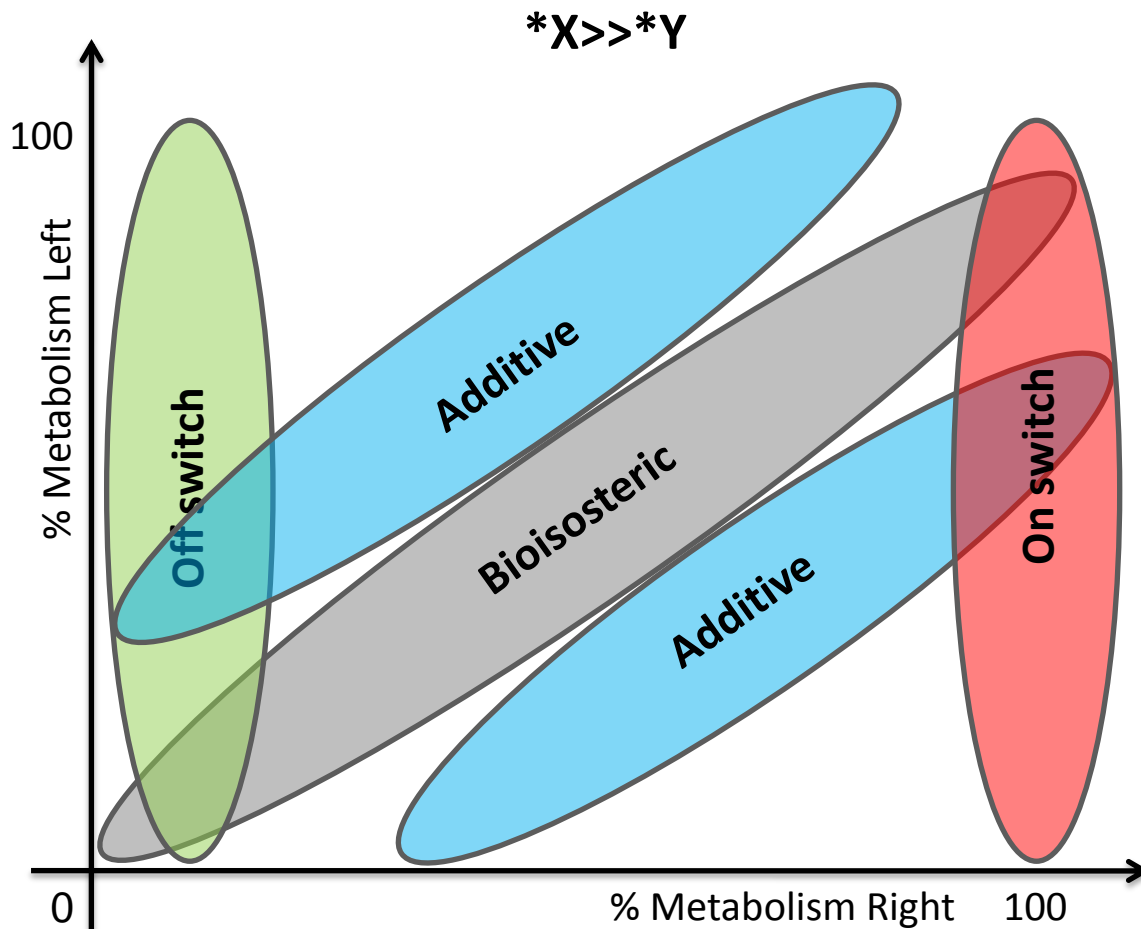


30 examples found | Mean Δ P: **5.51%**
Neutral Count Ratio: **96.55%**



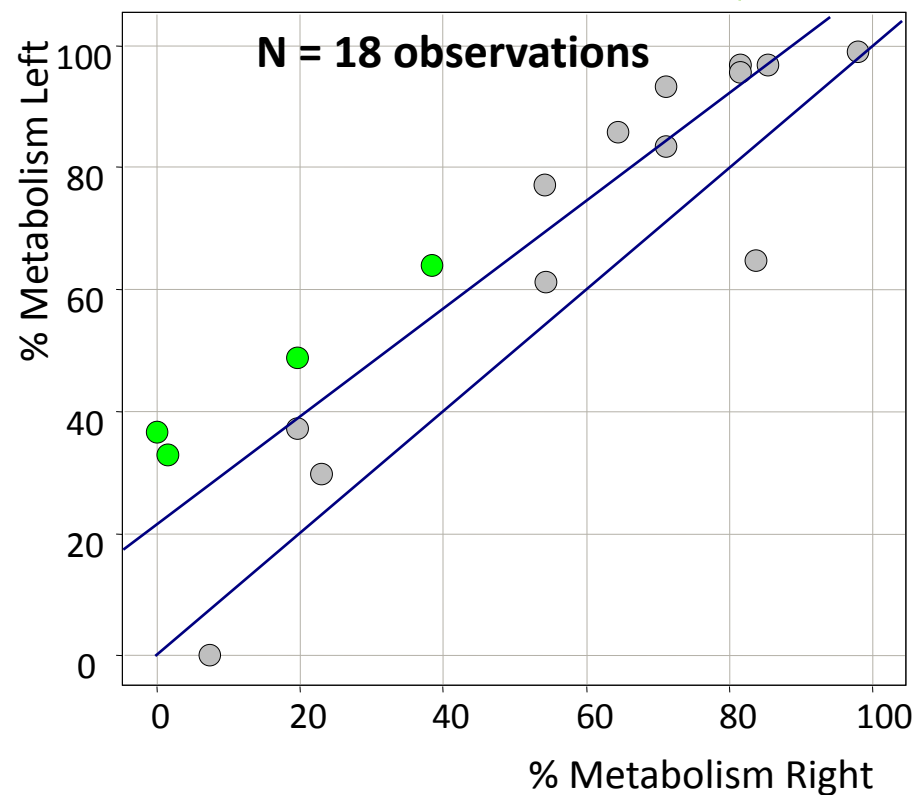
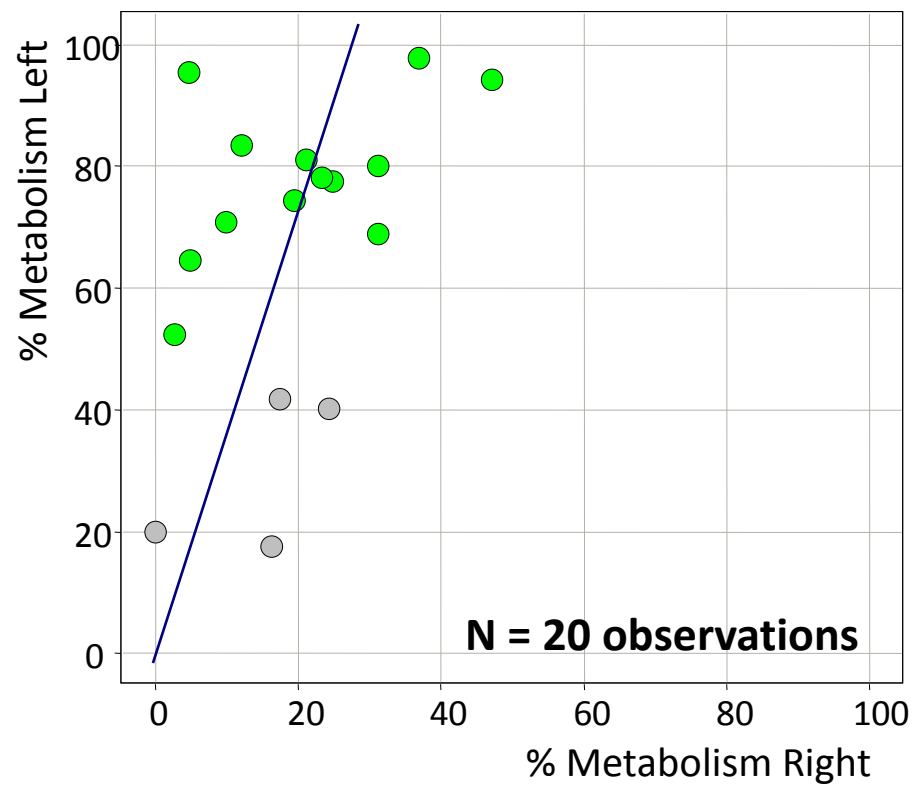
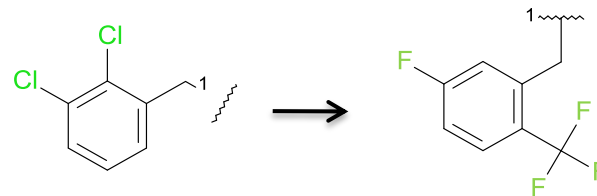
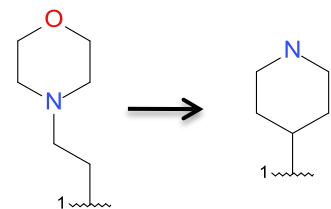
19 examples found | Mean Δ P: **2.8%**
Neutral Count Ratio: **100%**

Patterns in transformations



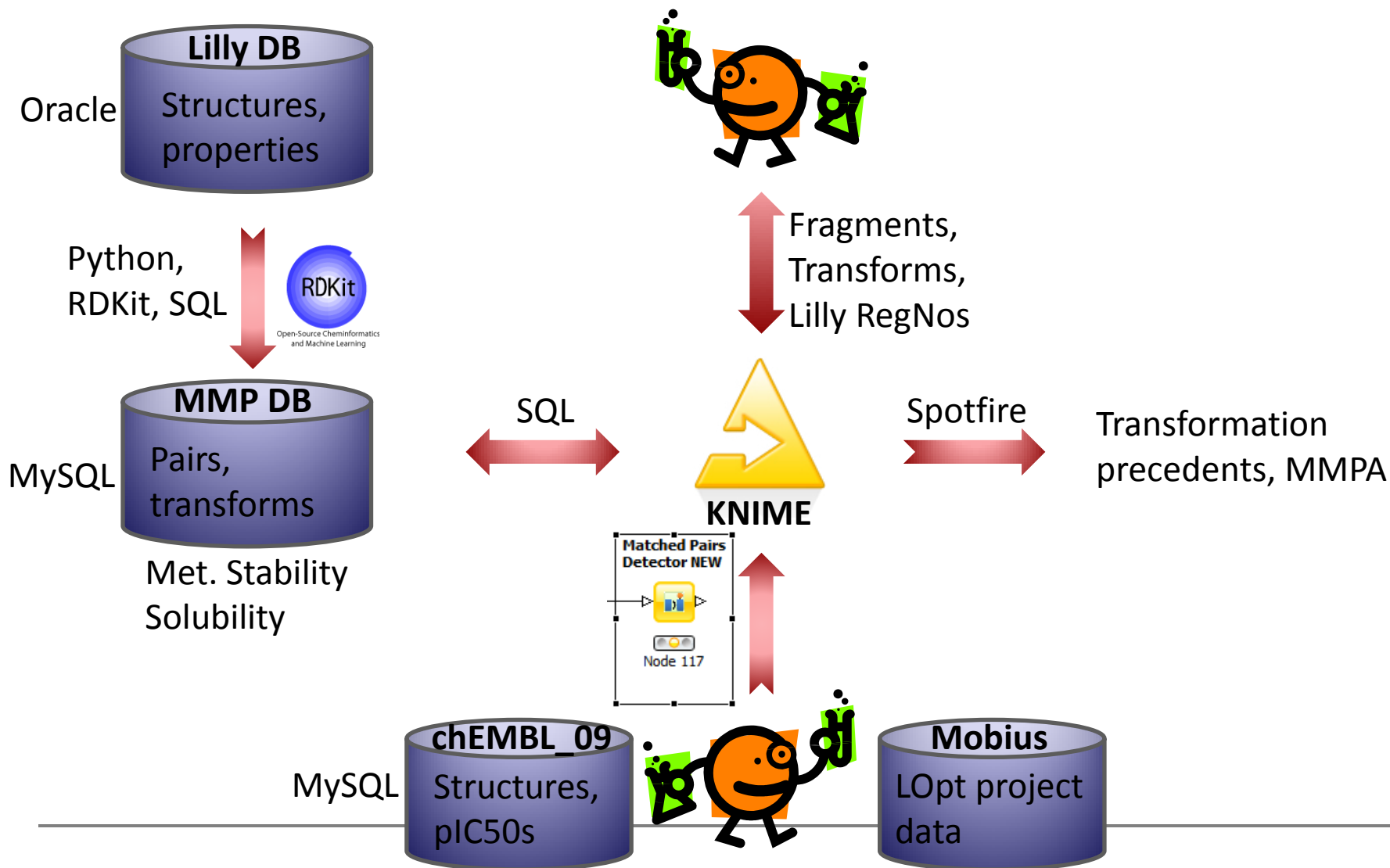
Keefer et al. "Extraction of tacit knowledge from large ADME data sets via pairwise analysis". *Bioorg. & Med. Chem., In Press*

Examples



Good
 Neutral
 Bad

Storing and searching for transformations



Matched Pairs Detector node

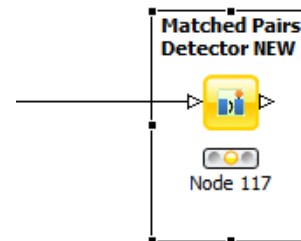
- Community Nodes
 - Erl Wood Cheminformatics
 - Activity Cliffs
 - Activity Cliffs Viewer
 - Calculators
 - Column Merger
 - Fingerprint Similarity
 - Virtual Screening Metrics
 - Convertors
 - Fingerprints Expander
 - Old Bit Vector To New Bit Vector
 - Docking
 - Docking Job Lister
 - Docking Job Retriever
 - Docking Job Submitter
 - IO
 - Chemical Reactions File Reader
 - Text Input
 - Multi-objective
 - Desirability
 - Pareto Ranking
 - RGroup Analysis
 - MCS Distance
 - MCS Matrix
 - Matched Pairs Detector**
 - Matched Pairs Finder
 - RGroup Efficiency
 - Reaction Generation
 - Reaction Generator
 - Reaction Vectors Database Reader
 - Reaction Vectors Database Writer
 - Viewers
 - Jmol Docking Pose Viewer
 - Jmol Viewer
 - Similarity Viewer
 - Vida Viewer

in RDKit format) and

ent molecule, IDs,
, ΔP , context,

(WinXP laptop, 1 CPU)

community contribution node
contributions-info



Dialog - 0:117 - Matched Pairs D...

File

Options | Flow Variables | Memory Policy

RDKit molecule column:

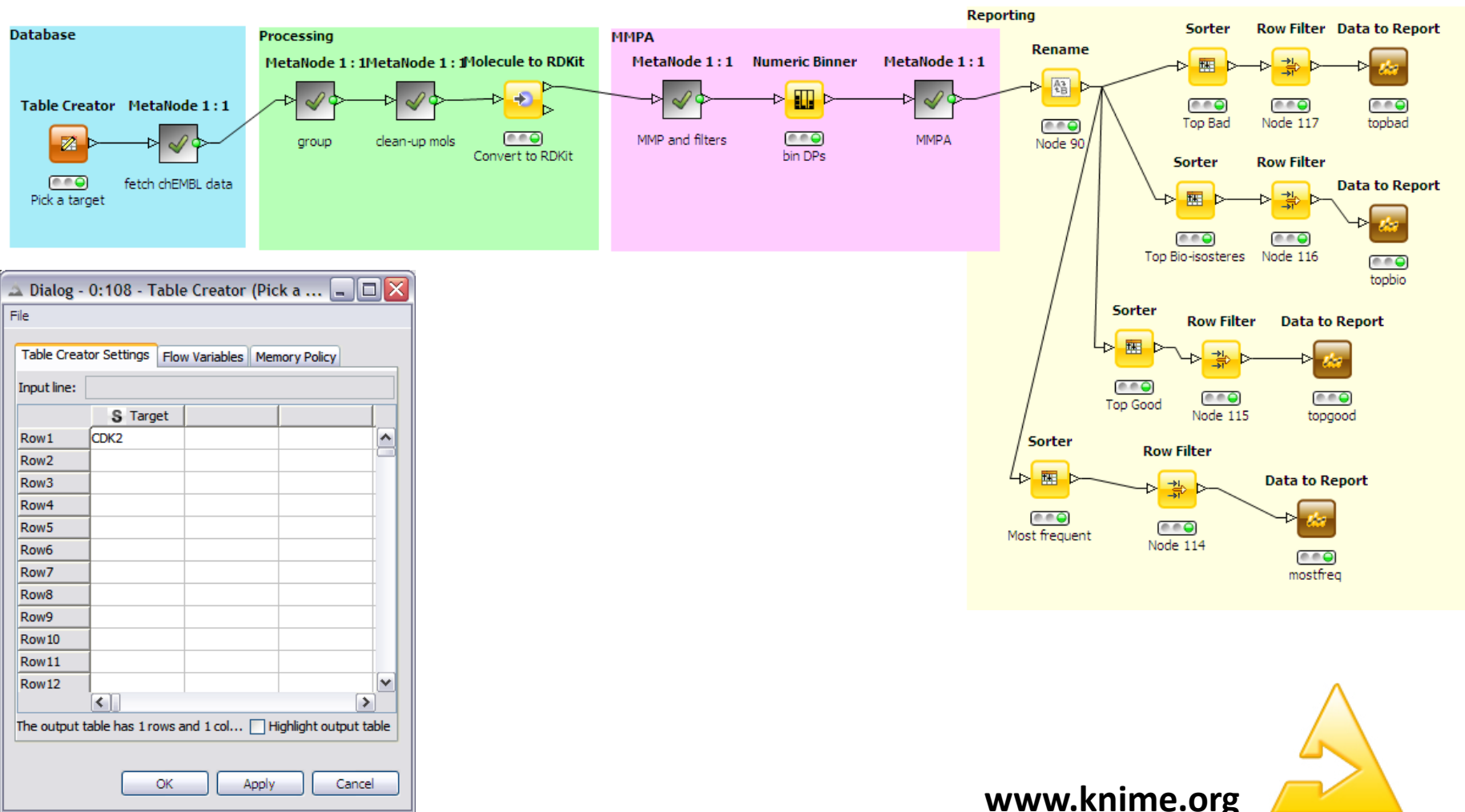
ID column:

Property column:

OK Apply Cancel

text	aw Fragment_L	aw Fragment_R	D Propert...	D Propert...	D Propert...	Trans...	MoReg...	D MolWe...
			0.56	5.49	6.05	24	1806	290.316
			-0.13	5.49	5.36	24	1806	290.316
			-0.49	5.49	5	3	1806	290.316
			-0.49	5.49	5	21	1806	290.316

Automated public domain MMPA

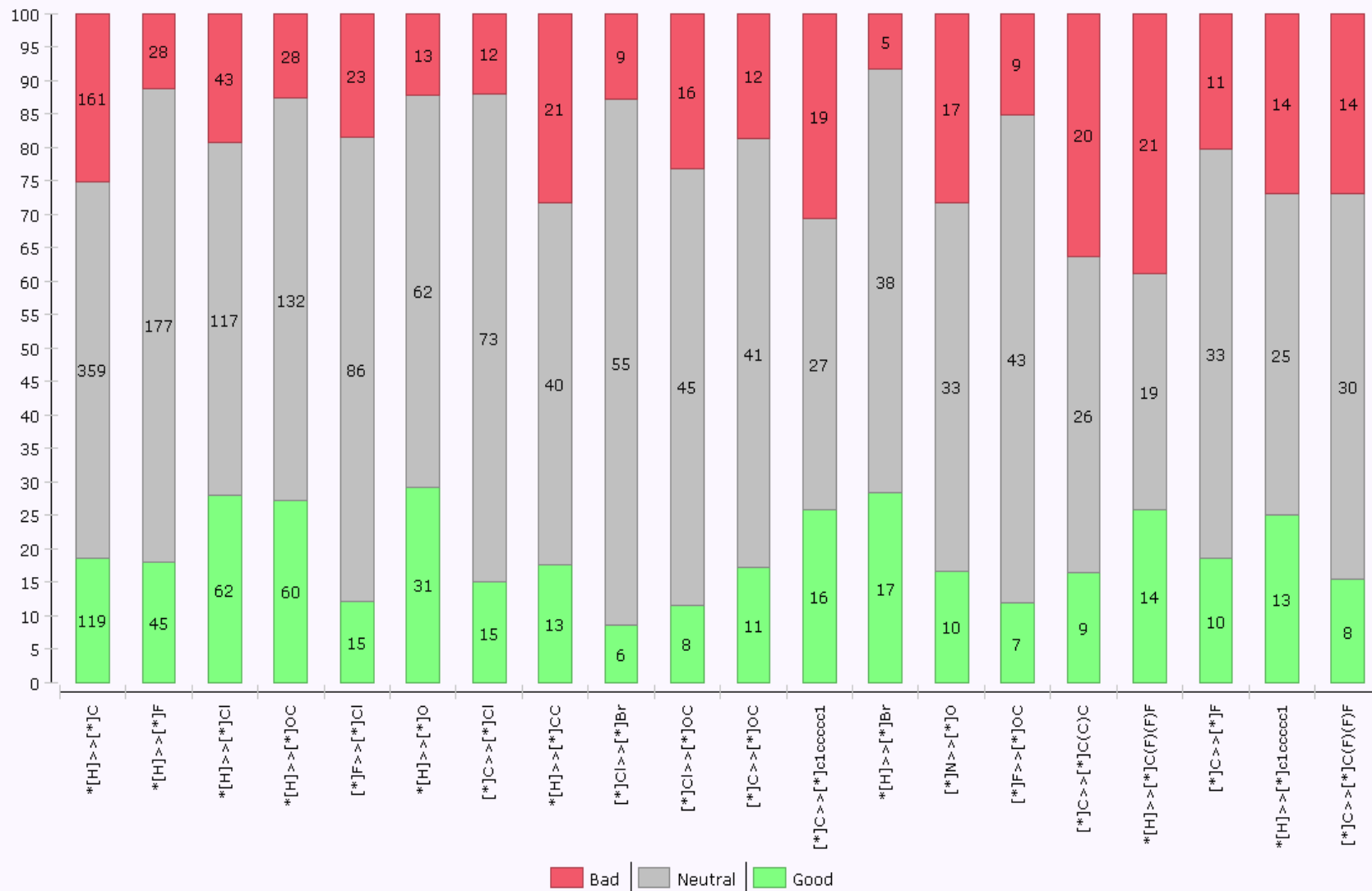


www.knime.org






MMP Analysis

Most Frequent Transformations



Generated report

Top biososteric transformations

Transformation	Count	Mean DP	StDev DP	BadCount	NeutralCount	GoodCount	BadCount%	NeutralCount%	GoodCount%
	10	0.04	0.21	0	10	0	0	100	0
	11	0.04	0.16	0	10	1	0	90.91	9.09
	17	-0.09	0.42	2	14	1	11.76	82.35	5.88

MMP look-up workflows

- Retrieve any given query substituent or transformation from a matched molecular pair database
 - Aqueous solubility and human met. stability
 - Real case scenario: Labile sites

- **Task:** Find replacements for these substituents that have proven successful in the past

Looking for alternative substituents

Structures - 2:86 - Chemical Sketcher

File Table 'default' - Rows: 4 Spec - Column: 1 Properties Flow Variables

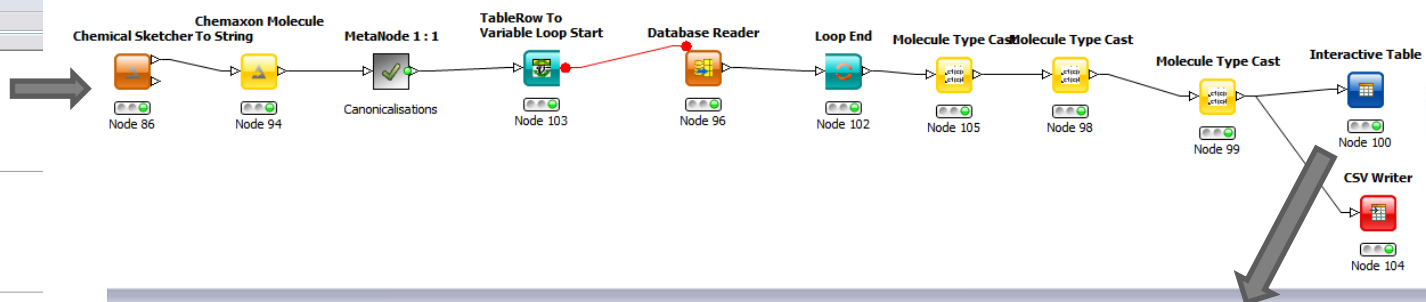
Row ID Molecule

1

2

3

4



Substituent search
MET. STABILITY

see smiles_per	see trans	see content	Stability	Stabil...	DP	DP_fm	deg1	deg2	deg3	Iteration
		<chem>COCA</chem> →	95.1	45.2	38.9	Good	4.97	3.852	-1.138	1
		<chem>COCA</chem> →	95.1	34.45	75.65	Good	4.97	4.525	-0.455	1
		<chem>COCA</chem> →	95.1	37.9	67.2	Good	4.97	4.628	-0.261	1
		<chem>COCA</chem> →	95.1	31.9	73.2	Good	4.97	4.953	-0.917	1
		<chem>COCA</chem> →	95.1	49.8	25.3	Good	4.97	4.573	-0.897	1
		<chem>COCA</chem> →	95.1	35.95	68.15	Good	4.97	4.528	-0.641	1

Summary

- MMP analysis
 - Provides data-driven, interpretable guidelines for LO
 - Can extract tacit knowledge from accumulated data beyond series, projects and departments
- Incorporation of contextual information is important
- KNIME MMP detector node
 - Allows for really fast, routine MMP analyses
 - Useful combination with public domain data
- MMP database integration with KNIME
 - Enables chemists to assess precedents of a proposed replacement
- Future plans
 - Apply the transformations prospectively and predict ΔP s

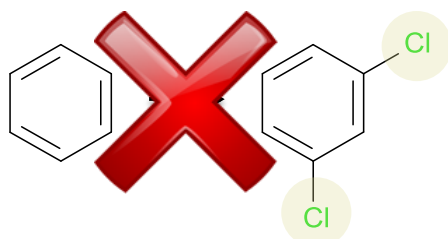
Acknowledgements

- Eli Lilly
 - Michel Bodkin, David Evans, Nikolas Fechner
- The University of Sheffield
 - Muhammad Alkarouri, Val Gillet, Visakan Kadiramanathan, Peter Willett
- GSK
 - Iain McLay, Chris Luscombe, Giampa Bravi, Nicola Richmond, Stephen Pickett
- KNIME

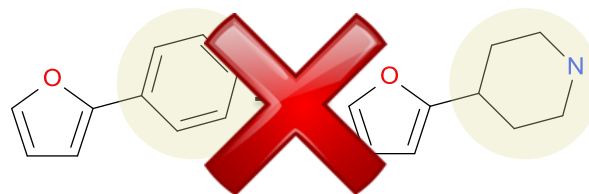
- All you for listening

Back-up slides

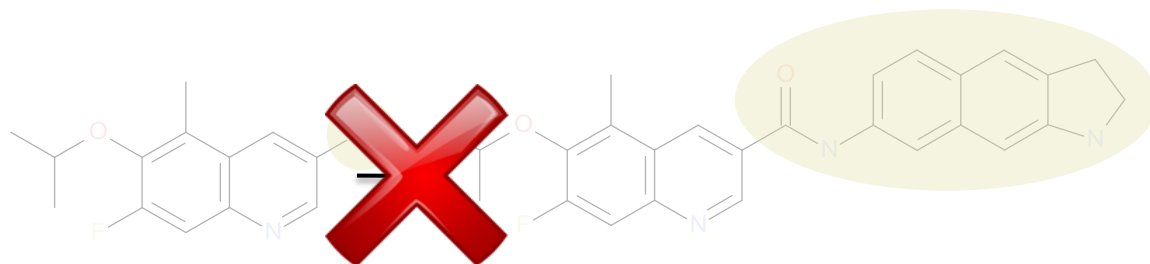
Non valid MMPs



Double change



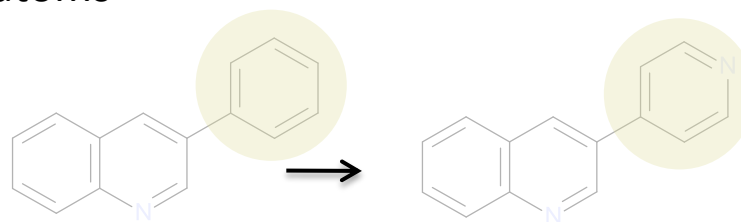
Change larger than shared part



Fragment larger than 15 heavy atoms



No breaking of rings



Phenyl to Pyridine transformation

Most frequent transformations

Rank	Transformation	# Examples	Rank	Transformation	# Examples
1	$\text{H}^\bullet \longrightarrow \text{---CH}_3$	5785	11	$\text{---CH}_3 \longrightarrow \text{---Cl}$	517
2	$\text{H}^\bullet \longrightarrow \text{---F}$	2530	12	$\text{---CH}_3 \longrightarrow \text{---F}$	458
3	$\text{H}^\bullet \longrightarrow \text{---Cl}$	1176	13	$\text{---CH}_3 \longrightarrow \text{---C(F)}_2$	453
4	$\text{H}^\bullet \longrightarrow \text{---OCH}_3$	886	14	$\text{---CH}_3 \longrightarrow \text{---OCH}_3$	438
5	$\text{---F} \longrightarrow \text{---Cl}$	617	15	$\text{H}^\bullet \longrightarrow \text{---C(CH}_3)_2$	412
6	$\text{H}^\bullet \longrightarrow \text{---CH}_2\text{CH}_3$	596	16	$\text{---F} \longrightarrow \text{H}_3\text{C---O}$	403
7	$\text{H}^\bullet \longrightarrow \text{---C(F)}_2$	594	17	$\text{---Cl} \longrightarrow \text{---C(F)}_2$	393
8		587	18	$\text{---CH}_3 \longrightarrow \text{---Cyclopropyl}$	390
9	$\text{---CH}_3 \longrightarrow \text{---C(CH}_3)_2$	542	19	$\text{H}^\bullet \longrightarrow \text{---C}\equiv\text{N}$	360
10	$\text{H}^\bullet \longrightarrow \text{---OH}$	522	20		349