Answers That Matter.

Matched Molecular Pairs *vs.* SVMs with RDKit and KNIME

David Evans and George Papadatos

Lilly Research Centre, Erl Wood Manor, Windlesham, UK

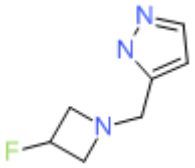
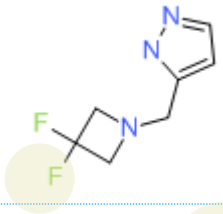
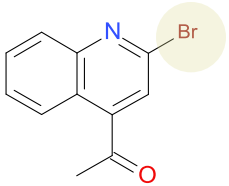
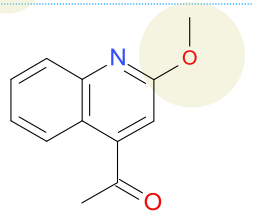
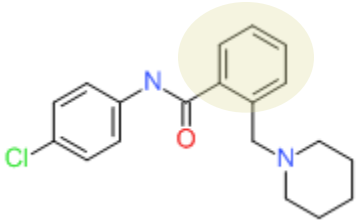
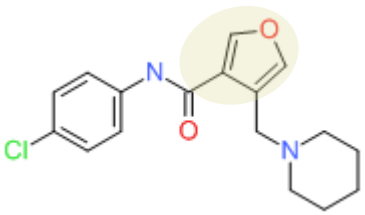
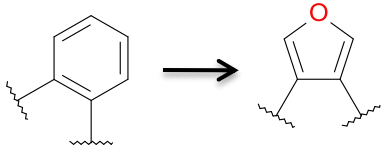
European Bioinformatics Institute, Hinxton, UK

4th October 2012

What is MMP analysis?

Answers That Matter.

The mining and statistical analysis of transformations and their impact on properties of interest (e.g. solubility or activity)

Left molecule	Right molecule	Transformation	ΔSolubility (mgml)
		H → F	-0.8
		Br → OCH_3	+1.2
			+2.4

Non-proprietary structures

Why predict with MMPs?

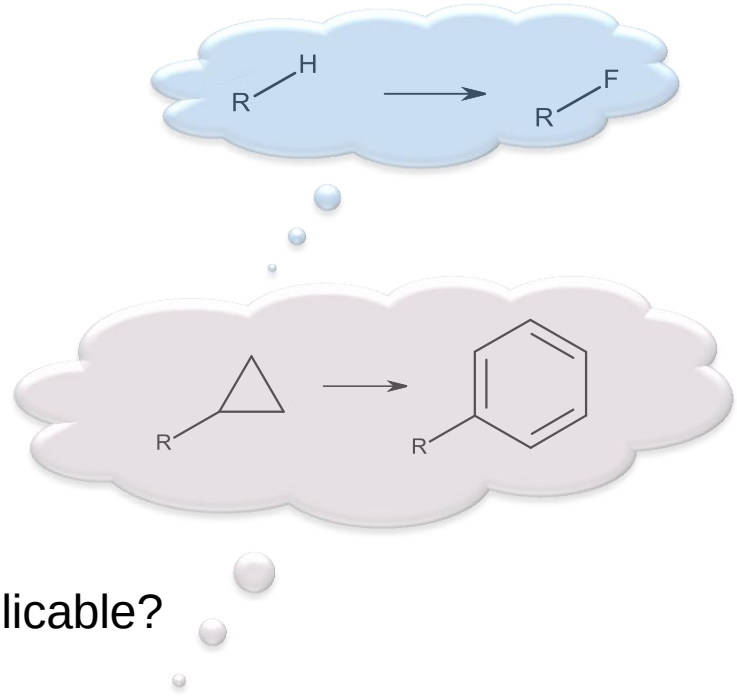
Answers That Matter.

Matched Pairs analysis provides interpretable trends

- If I swap group X for group Y what is likely to happen?

Driven by data rather than recollection

- Speak language of med-chem
- Deal with multiple objectives

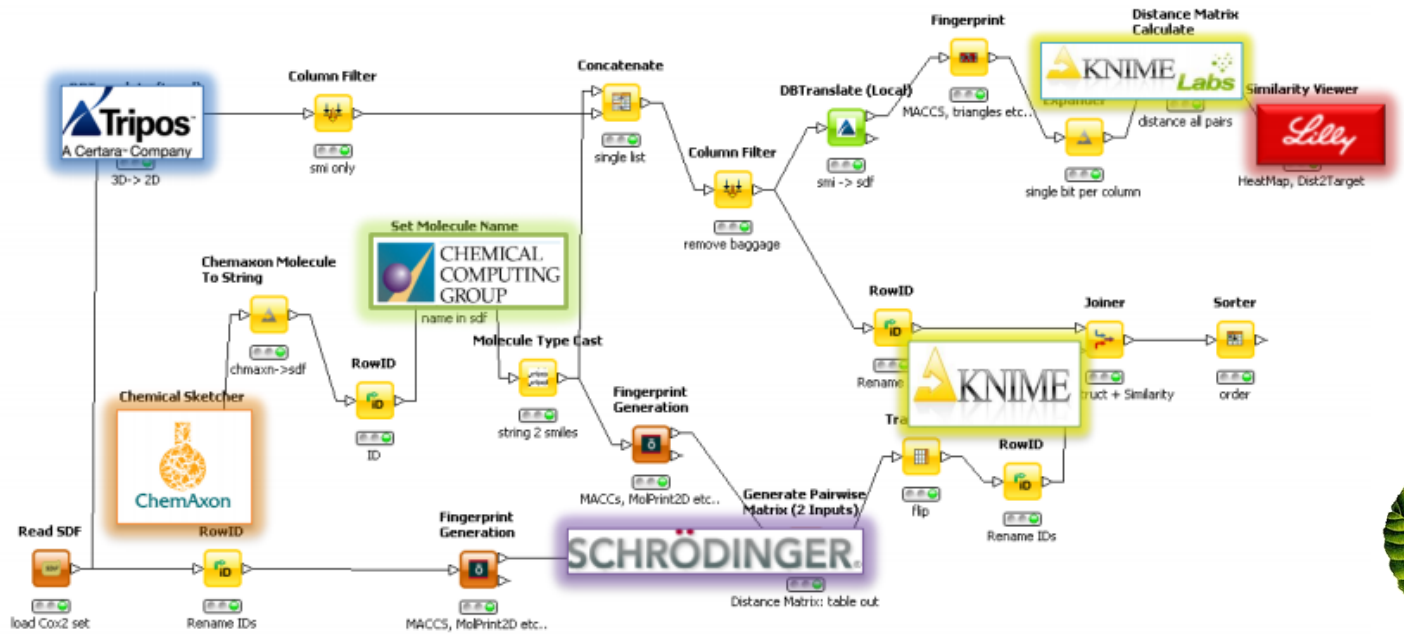


Can suggest molecules prospectively

- What transformations from the data base are applicable?
- What effects are likely?
- **Inverse QSAR approach**

KNIME

Answers That Matter.



Erl Wood Cheminformatics

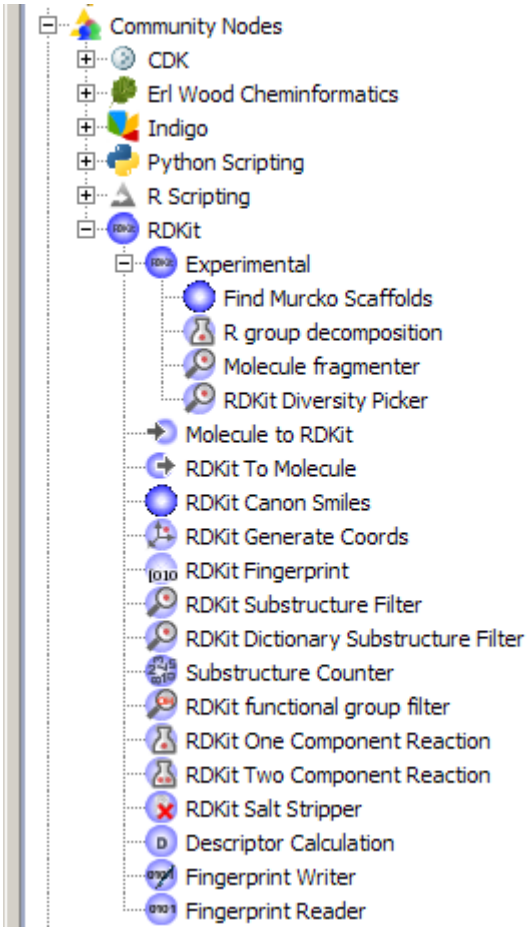
- Open Source workflow tool – desktop version is free
- But support is available and can integrate Open Source, commercial vendors + in-house code as nodes
- Have released many Erl Wood nodes to KNIME community site
  - http://tech.knime.org/community/erlwood

Workflow analysis in Lilly

Answers That Matter.

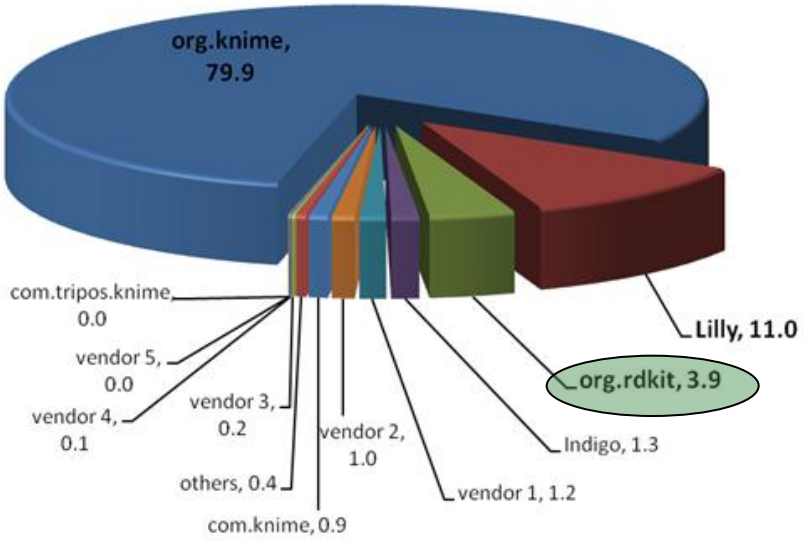
RDKit Nodes

KNIME 2.6.1 stable



- 1476 workflows from in-house server
- Count frequency of node usage

- RDKit most used external nodes (besides core KNIME)



David Thorner, James Lumley

How do we mine for MMPs?*

Answers That Matter.

(*in an automated and unsupervised way)

It used to be a slow and computationally expensive process...

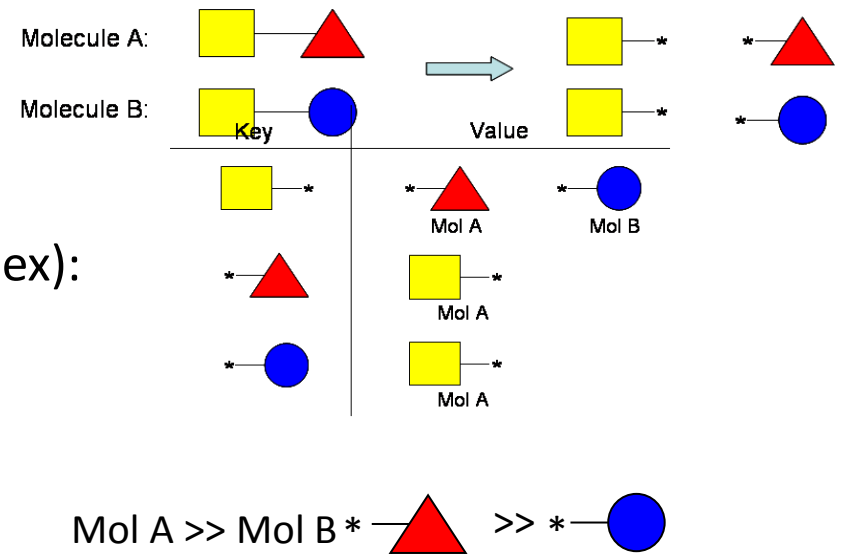
- Pair-wise maximum common substructure extraction – $O(N^2)$

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):

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.

Matched Molecular Pairs Detector Node

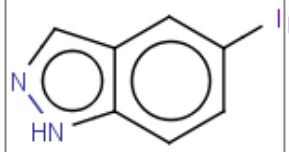
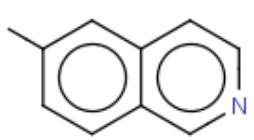
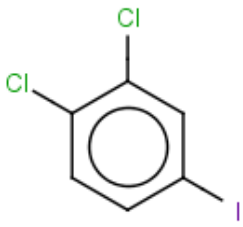
Answers That Matter

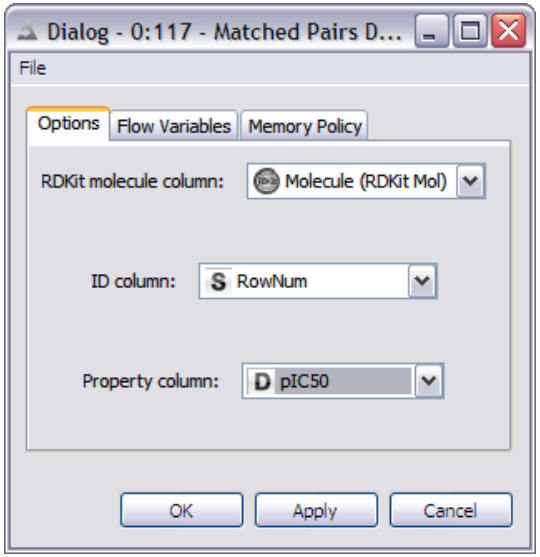
In: MolRegnos (IDs), structures (in RDKit format) and property values

Out: Matched pairs (left and right molecule, IDs, transformation, property values, ΔP, context, transformation atom count)

Implemented with RDKit Java API

Available as an **Erl Wood community contribution node**

Transformation_Arr[0]	Transformation_Arr[1]	BadCount	NeutralCount	GoodCount
		1	10	0
		1	26	3



KNIME vs. Python implementations

Lilly Answers That Matter.

Python

- Single, double and triple cuts
  - Will find R-group, linker and core transformations

```
SMART = "[*]!@!#!=[*]"                    # Cleavable bonds
pat = Chem.MolFromSmarts(SMART)
at_pairs = mol.GetSubstructMatches(pat) # a tuple of cleavable bonds as atom indices lists
at1, at2 = at_pairs[0]                    # Atom indices of the first pair
tmp = Chem.EditableMol(mol)               # Necessary to edit a molecule
tmp.RemoveBond(at1, at2)                  # Break the bond
tmp.AddAtom(Chem.Atom('Rb'))              # Introduce two dummy atoms in the molecule
tmp.AddAtom(Chem.Atom('Rb'))
tmp.AddBond(at1, numatoms, Chem.BondType.SINGLE) # Bond the Rb atoms to the new terminal atoms
tmp.AddBond(at2, numatoms + 1, Chem.BondType.SINGLE)
newmol = tmp.GetMol()
Chem.SanitizeMol(newmol)
#Use itertools.combinations(at_pairs, N) for N-cuts
```

Full Code not public

KNIME vs. Python implementations

Lilly Answers That Matter.

KNIME

- Uses Java API, ChemicalReaction module
- Single-cuts only
  - Will find R-group transformations only

```
//Reaction SMARTS that effectively cleaves all acyclic single bonds
String rea_smarts = "[*:1]!@!=!#[*:2]>>[*:1]-[*].[*:2]-[*]”;
ChemicalReaction rxn =
    ChemicalReaction.ReactionFromSmarts(rea_smarts);
mol = ((RDKitMolValue)mcell).readMoleculeValue();
ROMol_Vect rs = new ROMol_Vect(1);
rs.set(0, mol);
prods = rxn.runReactants(rs); // Magic happens here
```

Data-driven *de novo* design

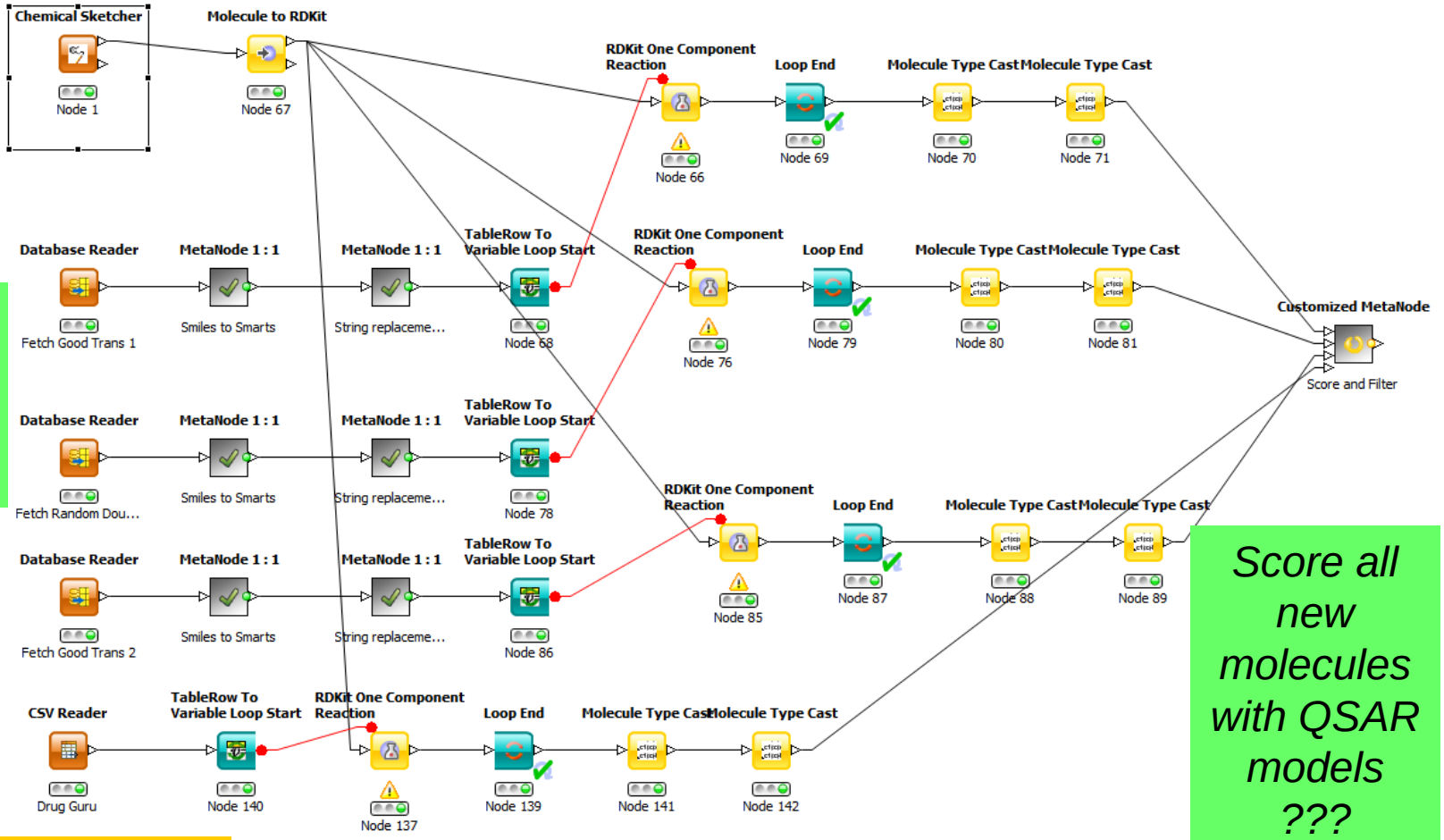
Answers That Matter.

Input starting point

Select effective Trans. from DB

Generic Trans

Python MMP code to build DB



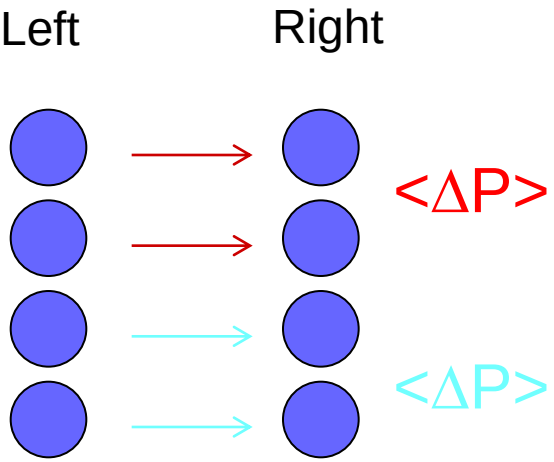
Score all new molecules with QSAR models ???

How to predict activity with MMPA?

Answers That Matter.

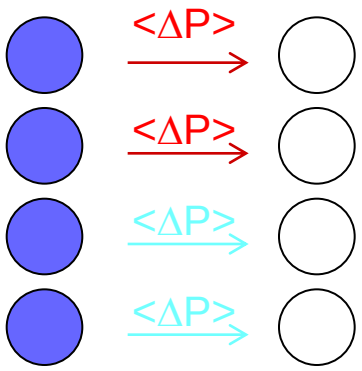
- Find all matched pairs
- Split pairs into train/test sets
- Group training set by transformation
  - For each transformation calculate $\langle\Delta P\rangle_{train}$
  - Keep if >10 examples

Train



- Test set
  - Predict activity of right-hand molecule
  - $P_{right} = P_{left} + \langle\Delta P\rangle_{train}$
  - Context information can be viewed

Test

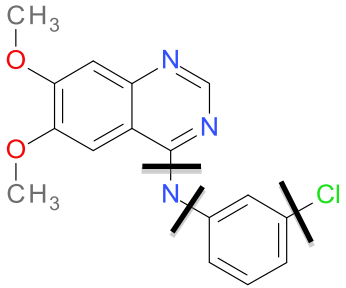


cf Griffen et al J Med Chem 2011 54 7739

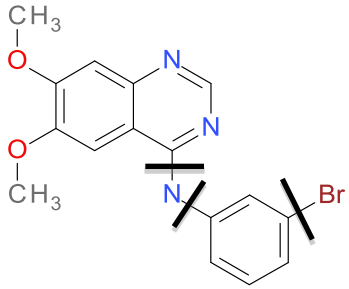
A note on context-dependency

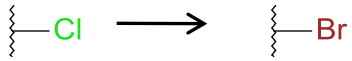
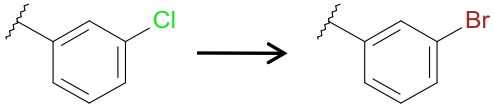
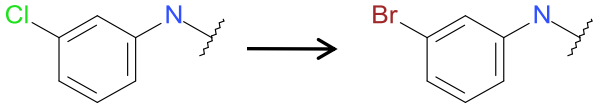
Answers That Matter.

left molecule



right molecule



	Heavy atoms involved	ΔpIC50 / ΔpK_i
	2	0.28
	14	0.28
	16	0.28

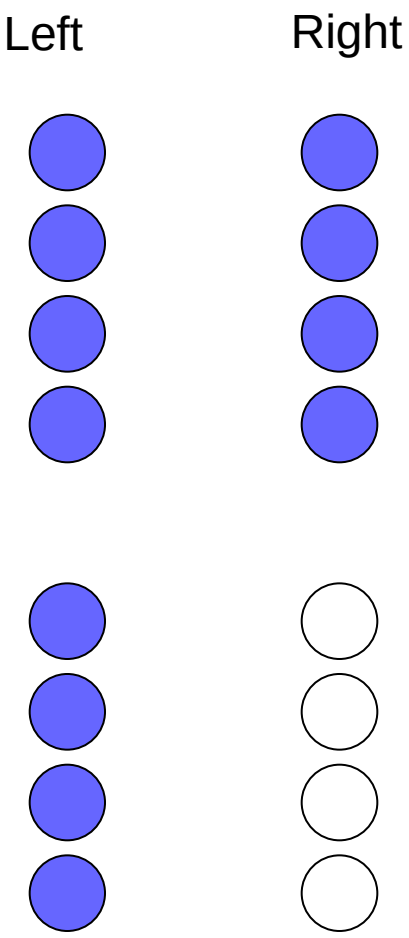
Papadatos et al. (2010). *J. Chem Inf. and Model.,* **50** (10), 1872-1886.

Non-proprietary structures

Comparison with SVM

Answers That Matter.

- Find all matched pairs
- Split pairs into train/test sets
- Train SVM on compounds in training pairs and left hand side of test pairs
- Test set
  - Predict activity of right-hand molecule with SVM model



SVM approach

Answers That Matter.

RDKit 2.0.0 Morgan fingerprints

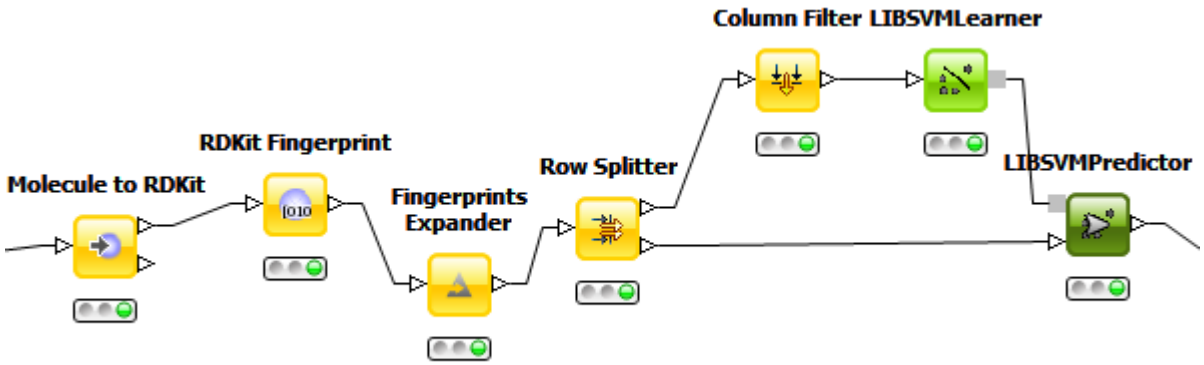
Radius 2, 1024 bits

LibSVM

Linear Kernel, epsilon regression

Initial validation on Sutherland data sets

J. Med. Chem. **47**, 5541–5554 (2004).



SVM validation

Answers That Matter.

$$R^2 = 1 - \frac{\langle (y_{pred} - y_{meas})^2 \rangle}{\langle (y_{meas} - \langle y_{meas} \rangle)^2 \rangle}$$

Data Set	R^2 test		
	RDKit-LibSVM	CoMFA-PLS*	HQSAR-PLS*
ACE	0.46	0.49	0.30
AchE	0.59	0.47	0.37
BZR	0.18	0.00	0.17
COX2	0.22	0.29	0.27
DHFR	0.58	0.59	0.63
GPB	0.33	0.42	0.58
THER	0.35	0.54	0.53
THR	0.21	0.63	-0.25

* Data from *J. Med. Chem.* **47**, 5541–5554 (2004).

ChEMBLdb data sets

Answers That Matter.

ChEMBLdb Kinase and Protease inhibitors

- Using chEMBL_14 (current version)
- Med. chem. friendly compounds, parent structure, confidence score >7 exact IC50 or K_i values only (converted to pIC50/pK_i)
- Multiple measurements for a compound and target were averaged
  - If Standard deviation > 1, measurements were discarded
- 10 most populated sets in each class

Kinase Name	P Acc	N
Vascular endothelial growth factor receptor 2	P35968	1341
MAP kinase p38 alpha	Q16539	1065
Epidermal growth factor receptor erbB1	P00533	724
Tyrosine-protein kinase SRC	P12931	624
Receptor protein-tyrosine kinase erbB-2	P04626	538
Hepatocyte growth factor receptor	P08581	529
Serine/threonine-protein kinase AKT	P31749	479
Tyrosine-protein kinase LCK	P06239	396
Serine/threonine-protein kinase Chk1	O14757	395
Glycogen synthase kinase-3 beta	P49841	392

Protease Name	P Acc	N
Coagulation factor X	P00742	1525
Dipeptidyl peptidase IV	P27487	1406
Thrombin	P00734	1097
Cathepsin S	P25774	925
Cathepsin K	P43235	895
Caspase-3	P42574	883
Epoxide hydratase	P34913	685
Matrix metalloproteinase-2	P08253	668
Matrix metalloproteinase 13	P45452	635
Beta-secretase 1	P56817	586

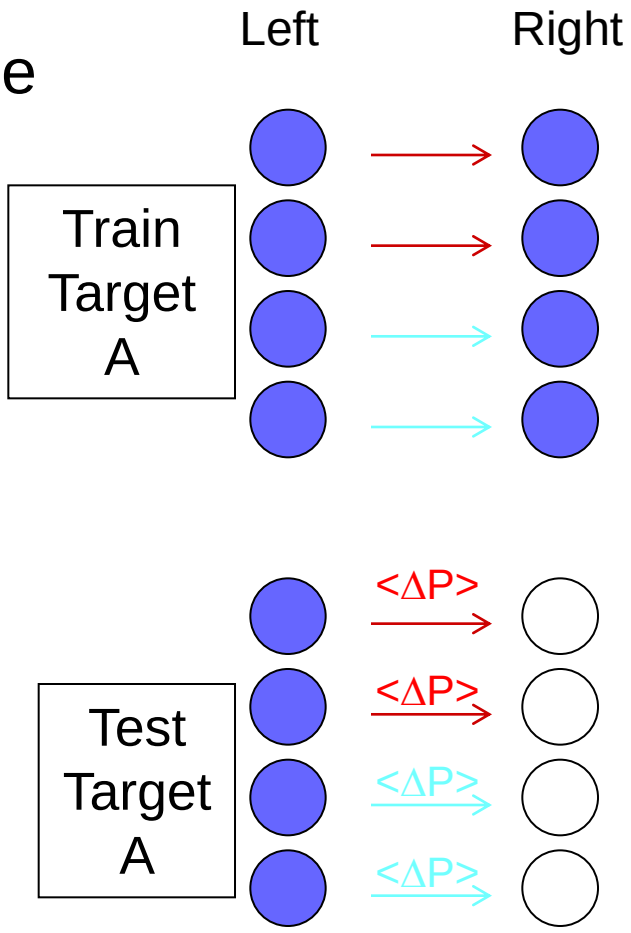
Results: Within data set

Answers That Matter.

- 50/50 test/train split within each target data set
- MMPA and SVM performance comparable

Kinases	MMPA R^2 test	SVM R^2 test
CHEK1	0.37	0.42
EGFR	0.76	0.8
ERBB2	0.86	0.88
LCK	0.33	0.54
MET	0.45	0.39
SRC	0.73	0.8
AKT1	0.4	0.68
KDR	0.46	0.42
GSK3B	0.5	0.19
MAPK14	0.62	0.52

Proteases	MMPA R^2 test	SVM R^2 test
F2	0.62	0.64
F10	0.77	0.71
MMP2	0.46	0.6
CTSS	0.45	0.69
DPP4	0.53	0.51
EPHX2	0.56	0.64
CASP3	0.75	0.68
CTSK	0.73	0.63
MMP13	0.59	0.58
BACE1	0.53	0.59



Across data sets: Same family

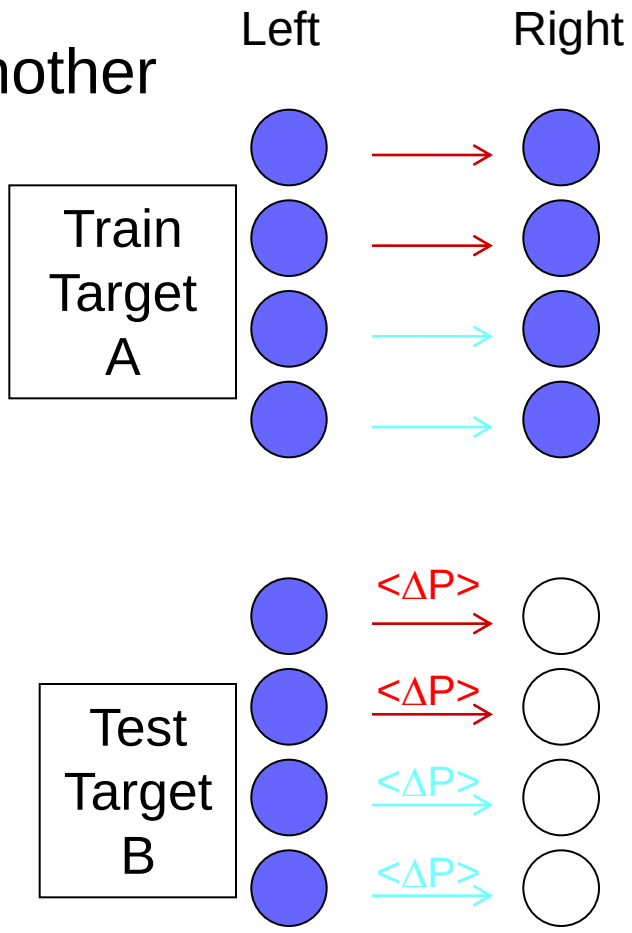
Answers That Matter.

- Find most similar data set in same family – average Morgan similarity across ligands
- Train models on one target, predict on another

- MMPA still performs reasonably well
- SVM fails

Kinases	MMPA R^2 test	SVM R^2 test
CHEK1	0.11	-1.93
EGFR	0.77	0.46
ERBB2	0.86	-0.06
LCK	0.4	-0.34
MET	0.46	-3.03
SRC	0.74	-0.44
AKT1	-0.05	-2.65
KDR	0.27	-3.09
GSK3B	0.19	-6.36
MAPK14	0.3	-3.72

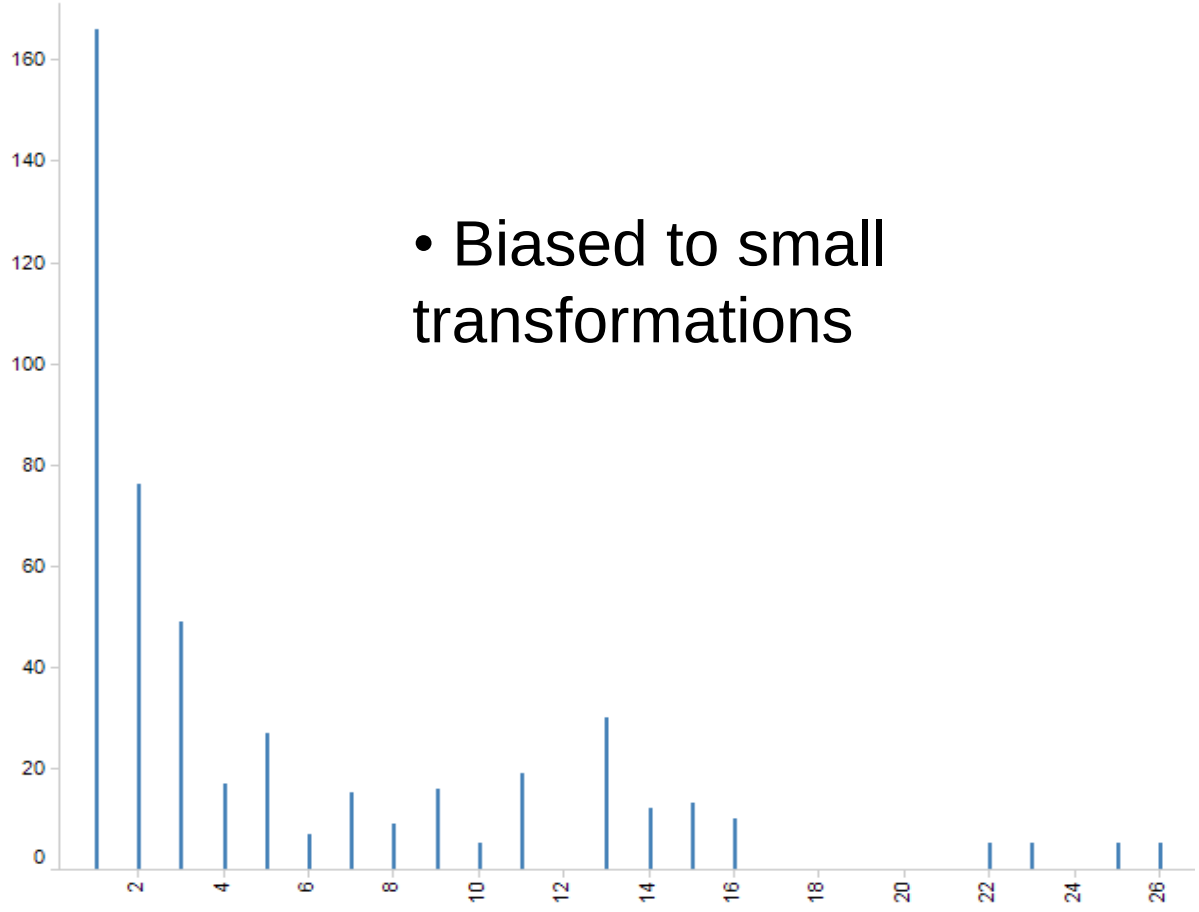
Proteases	MMPA R^2 test	SVM R^2 test
F2	0.73	-2.86
F10	0.43	0.07
MMP2	0.48	-2.41
CTSS	0.16	-3.46
DPP4	-0.12	-8.88
EPHX2	0.58	-1.58
CASP3	0.75	-0.24
CTSK	0.57	-3.31
MMP13	0.32	-3.7
BACE1	0.73	-2.86



What sort of transformations?

Answers That Matter.

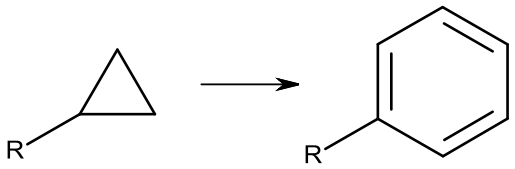
Within Kinase



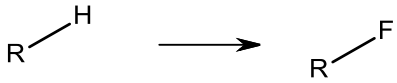
- Biased to small transformations

N_{atoms}

Number of heavy atoms involved in transformation



$N_{atoms} = 9$



$N_{atoms} = 1$

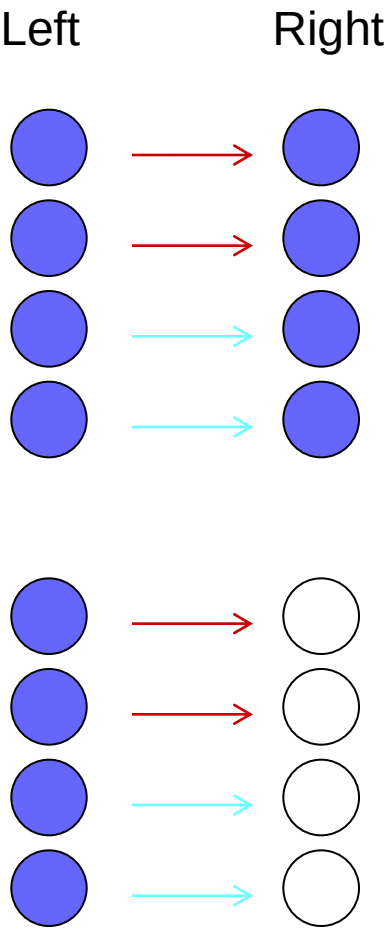
Remove Small Transformations

Answers That Matter.

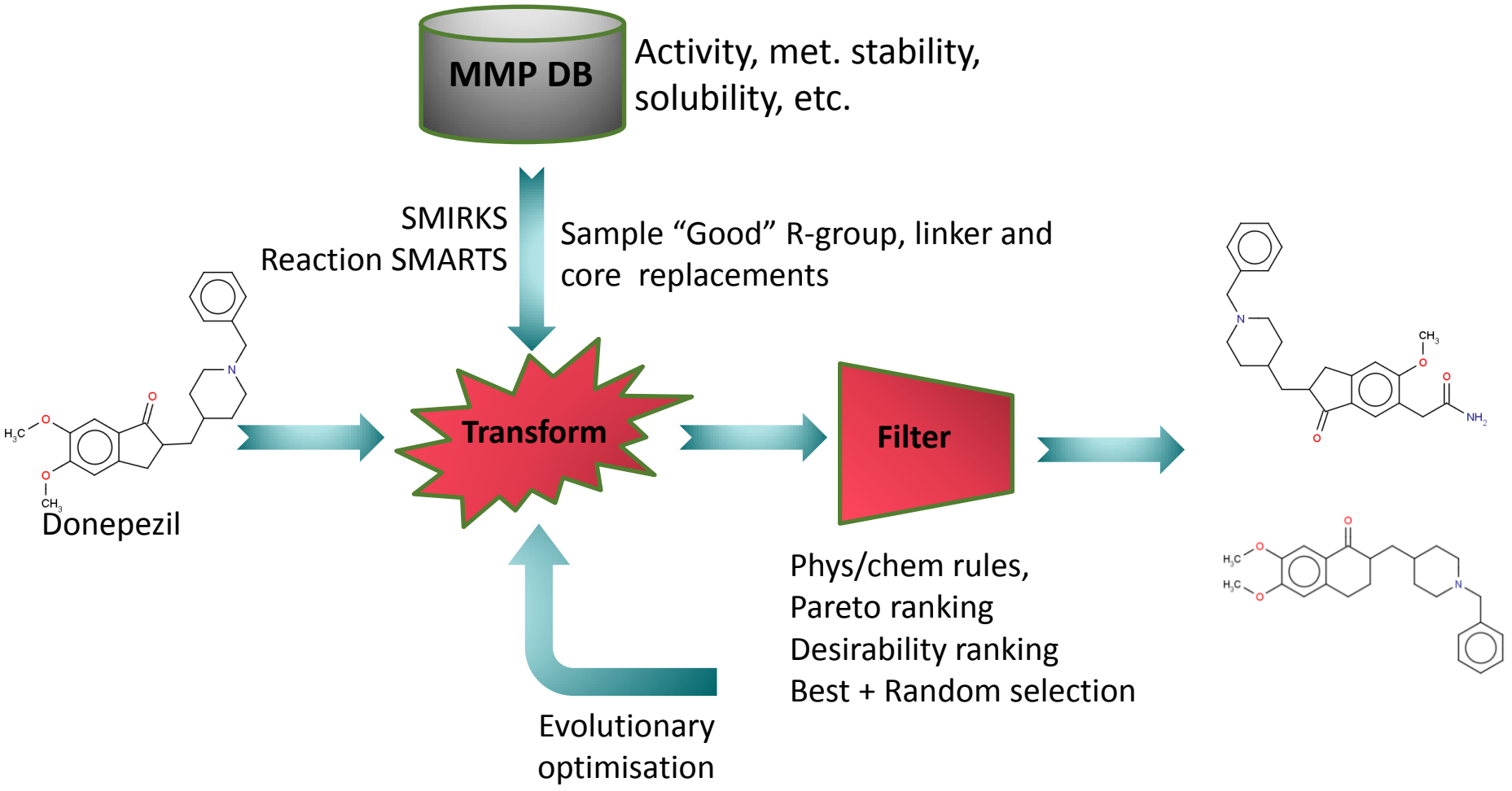
- Remove transformations with N_{atoms} <3 from test set
- Performance deteriorates
- Still can be useful even across targets

Within Kinase Sets	MMPA R^2 test Filtered	MMPA R^2 test Full
CHEK1		0.37
EGFR	0.6	0.76
ERBB2	0.83	0.86
LCK	-0.32	0.33
MET		0.45
SRC	0.82	0.73
AKT1	0.18	0.4
KDR	0.49	0.46
GSK3B	-0.84	0.5
MAPK14	0.67	0.62

Across Kinase Sets	MMPA R^2 test Filtered	MMPA R^2 test Full
CHEK1		0.11
EGFR	-0.26	0.77
ERBB2	0.59	0.86
LCK	0.11	0.4
MET		0.46
SRC	0.66	0.74
AKT1	-1.79	-0.05
KDR	0.13	0.27
GSK3B	0.6	0.19
MAPK14	0.44	0.3



Next stop: *de novo* design



Stewart et al. (2006). *Bioorg. & Med. Chem.,* **14** (20), 7011-7022.

Conclusions and Questions

Answers That Matter.

- Prediction performance with MMPA comparable to SVM within data set
- RDKit + LibSVM comparable to published QSAR benchmarks
- Can use transformations from other targets to maintain or enhance activity
    - Combine with established work on solubility, metabolism for multi-objective transformations

- http://tech.knime.org/community/erlwood

Acknowledgements

David Thorner

James Lumley

Hina Patel

Nikolas Fechner

Michael Bodkin

KNIME, ChEMBL + RDKit !