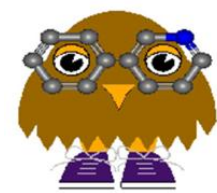


WizePairZ: Auto-Curation of Matched Molecular Pairs

David Wood

Ed Griffen, Steve St-Gallay, Dan Warner

ICCS 2011, Noordwijkerhout

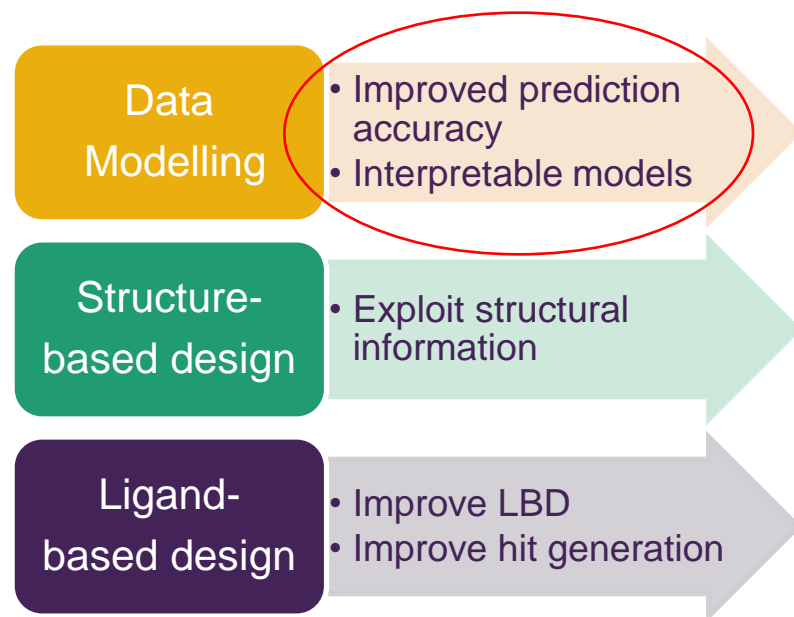


Overview

- Predictive Chemistry Programme at AstraZeneca
- Matched Molecular Pair Analysis (MMPA)
 - QSAR versus Inverse QSAR
- WizePairZ
 - Data Warehouses
 - Mining for MMPs
 - Finding Useful Molecular Transformation Rules
 - The Rule Database
- Applying WizePairZ Rules to New Compounds
 - Retrospective Gluco-Kinase Activators (GKA) Project Example
- Summary

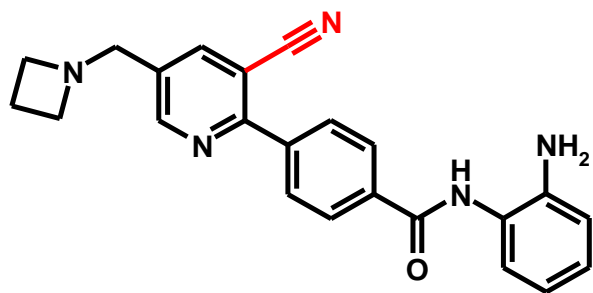
Finding Better Compounds Faster

- **Predictive Chemistry Programme @ AstraZeneca**
 - External collaborations
 - Internal projects
- **Aims:**
 - Enhance design quality and efficiency
 - Improve prediction accuracy
 - Enhance and exploit Chemistry Intelligence platform
- **Tools:**

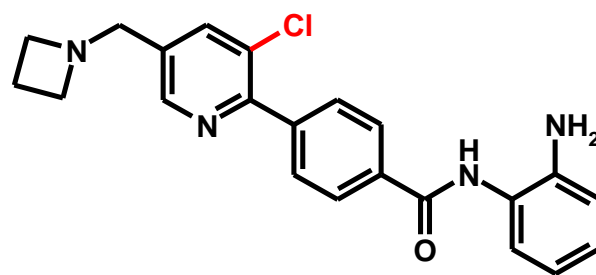


Matched Molecular Pairs

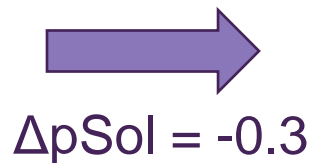
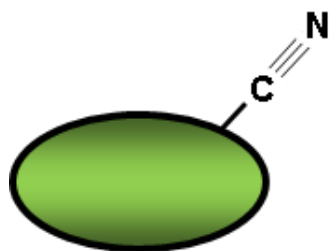
“Molecules that differ only by a particular, well-defined, structural transformation” Leach et al. ¹



pSol = 3.2

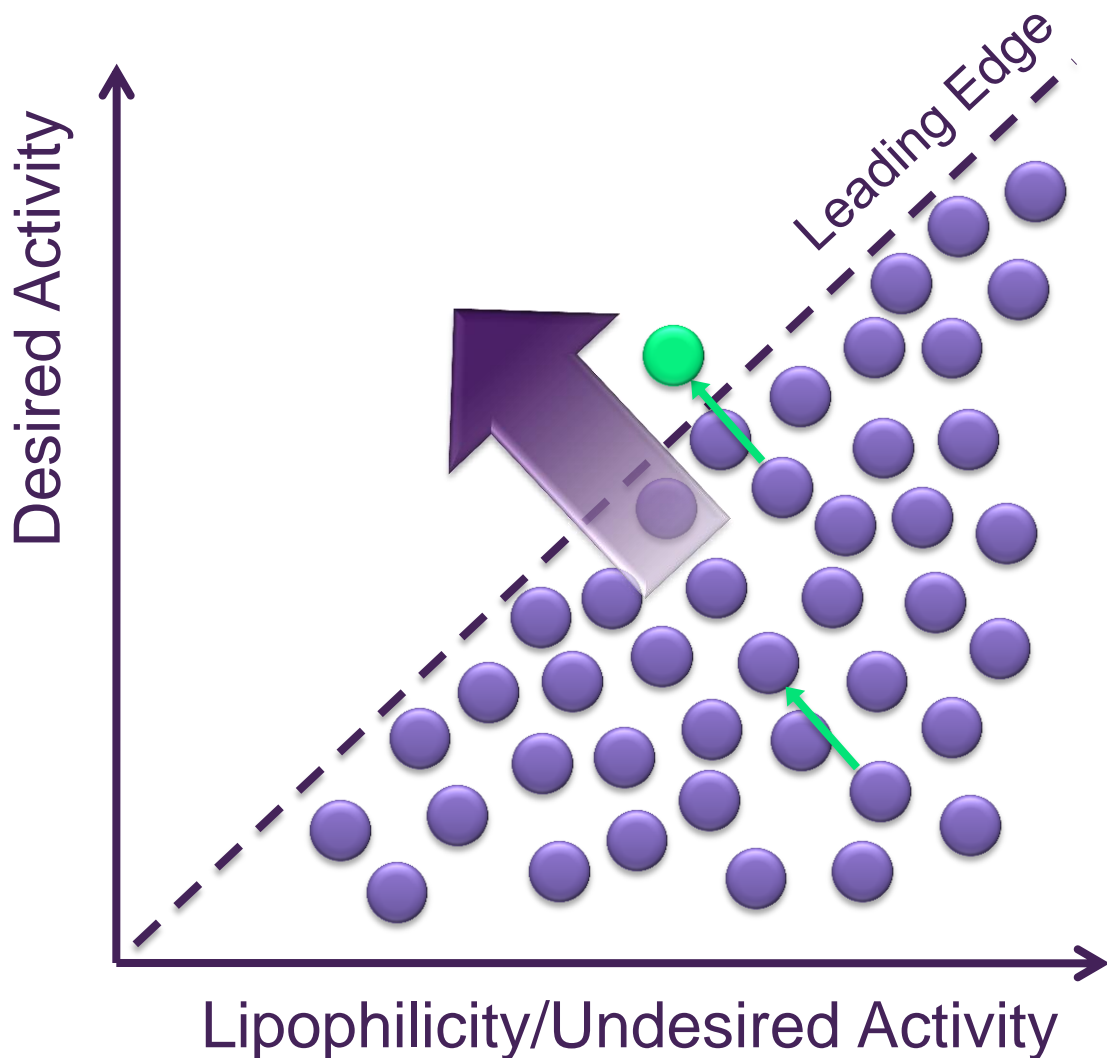


pSol = 2.9



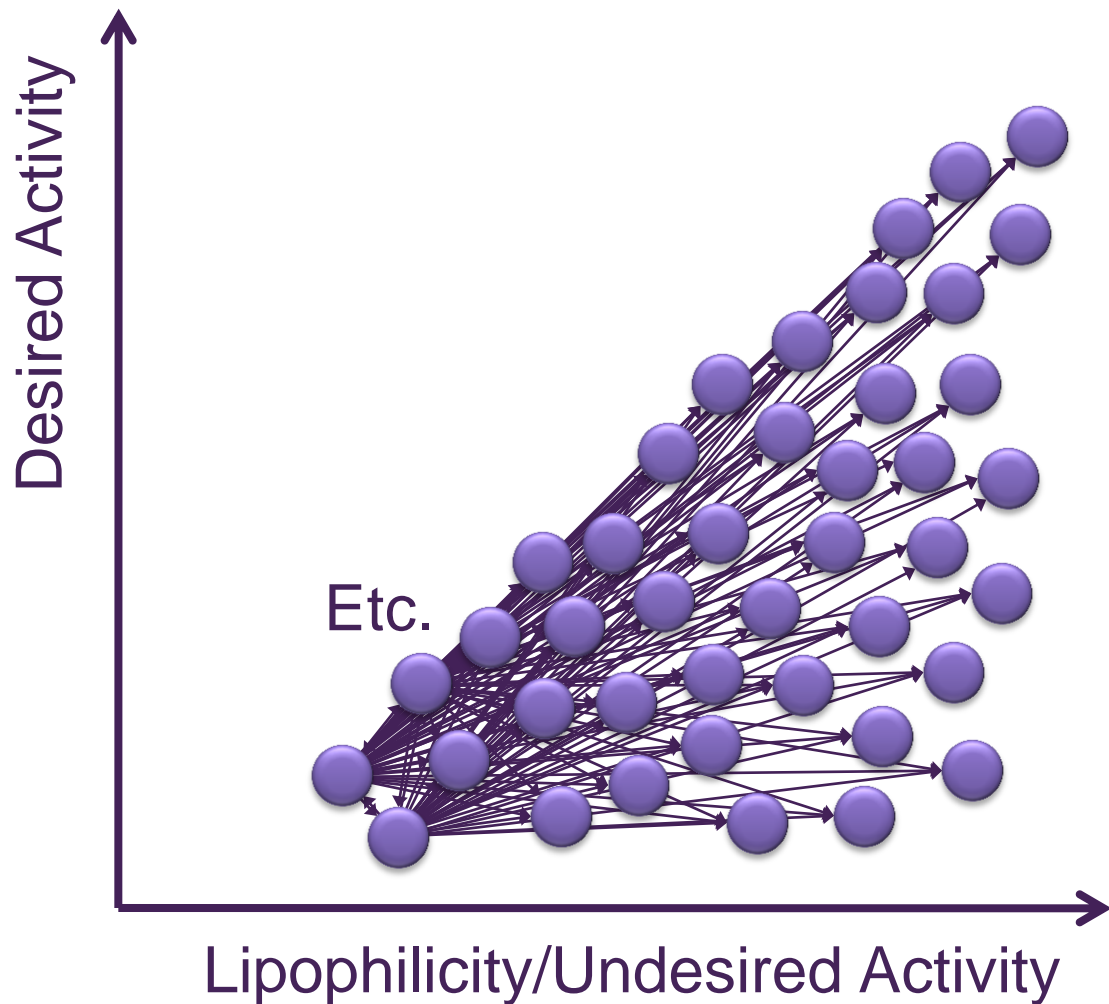
1. Leach et al. *J. Med. Chem.* 2006, 49, p6672.

Typical Problem for Drug Discovery Projects



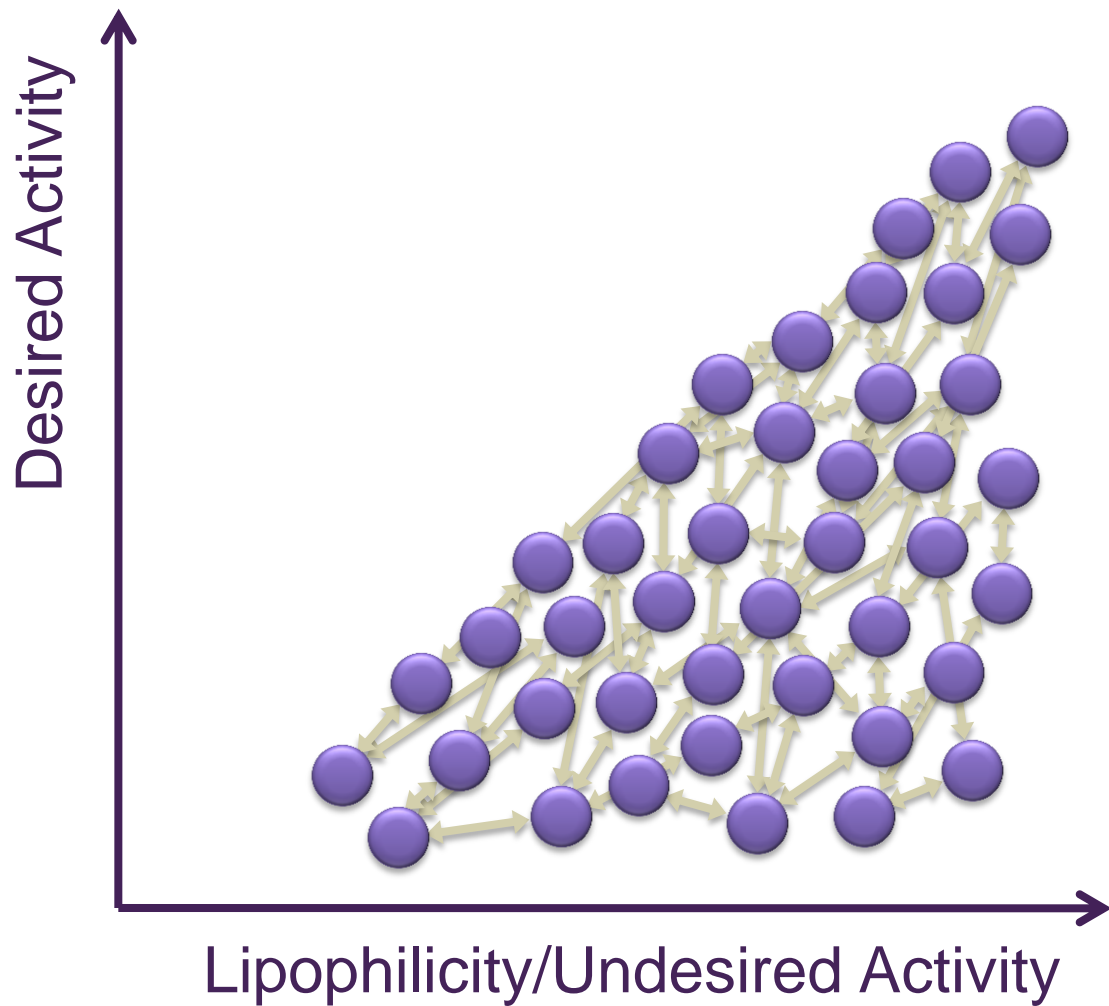
- Desired activities can correlate with undesired activities through lipophilicity
- Achieving a breakthrough requires pushing the 'leading edge'
- Molecular properties governed by structure, not just lipophilicity
 - Transformations orthogonal to the leading edge
- Exploit the structural variations to push the leading edge

Matched Molecular Pair Analysis



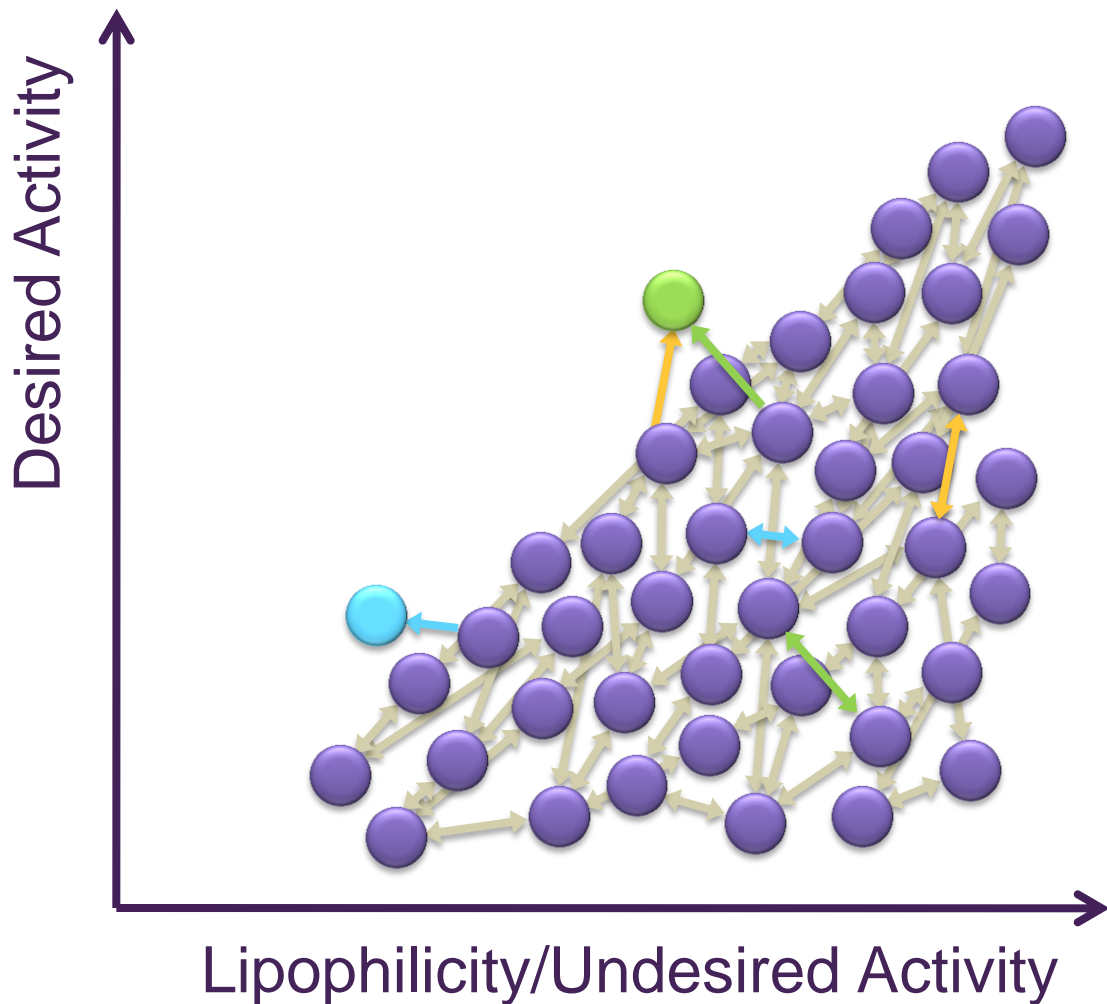
- Find MMPs for each compound in the set

Matched Molecular Pair Analysis



- Find MMPs for each compound in the set

Matched Molecular Pair Analysis



- Find MMPs for each compound in the set
- Apply the transformations to the existing compounds
 - Do they push the leading edge?
- Multiple transforms suggesting the same compound

QSAR versus Inverse QSAR

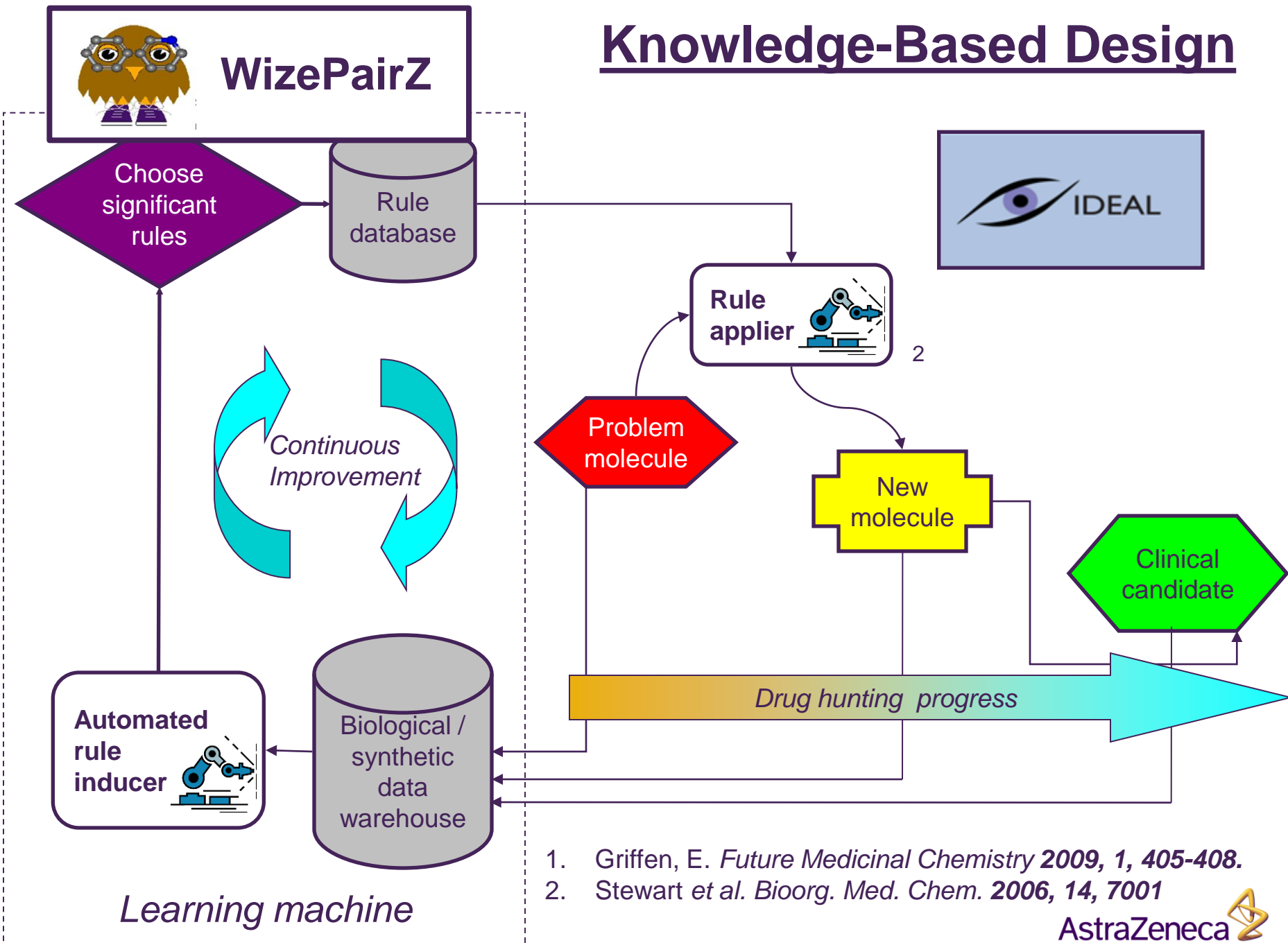
QSAR

- Compounds already proposed
- Prioritize *virtual compound sets* for synthesis
- Usually general SAR trends identified
- Models can be abstract and lack clear interpretation
- Usually applicable to all chemistries represented in the training set

MMPA

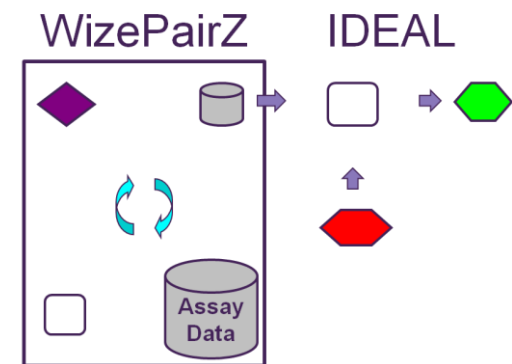
- Proposes new compounds
- Fix specific issues on *single compounds*
- SAR fine structure explored
- Clear link between transformations and the underlying data
- Limited to matched pair transformations that have previously been seen

Knowledge-Based Design



1. Griffen, E. *Future Medicinal Chemistry* 2009, 1, 405-408.
2. Stewart et al. *Bioorg. Med. Chem.* 2006, 14, 7001

AstraZeneca Data Warehouse



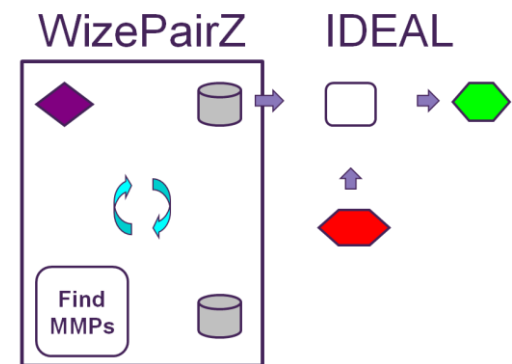
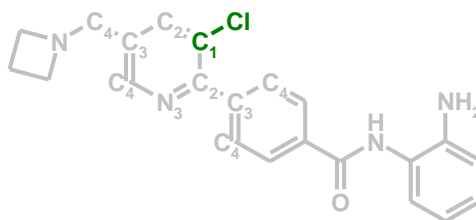
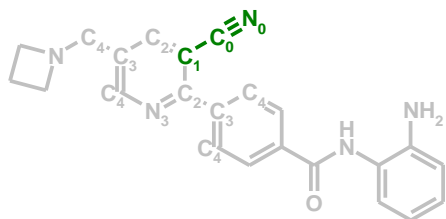
Global Datasets

- >100,000 measurements
 - LogD
- 10,000s of measurements
 - Solubility, hERG
- 1,000s of measurements
 - CYP Inhibition,
Other Cardiac Channels

Local Datasets

- Hepatocyte Clearance
 - Site based assays with 1,000s of measurements
- Potency
 - 100s of assays with up to 1,000s of data points

Mining for Matched Molecular Pairs



■ WizePairZ 1:

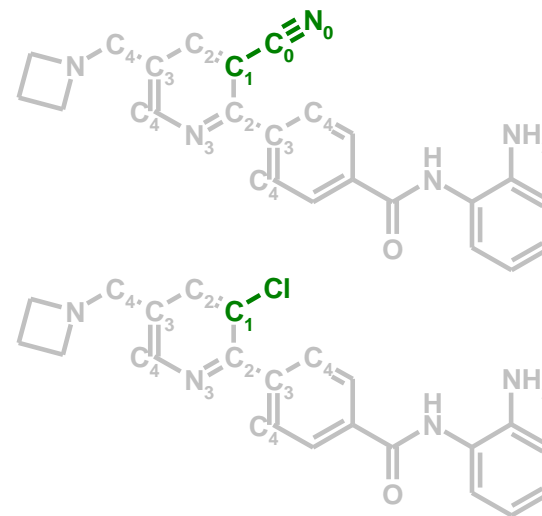
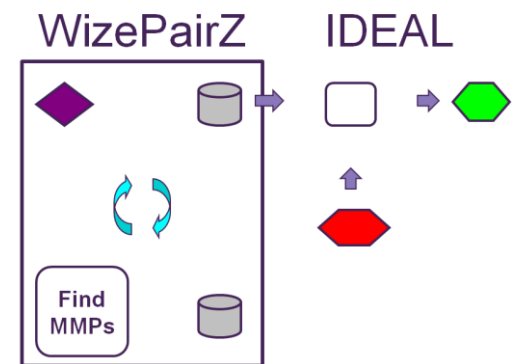
- Pairs found with Maximum Common Subgraph searches
 - Computationally expensive
- MCS results stored permanently in a database
- Transformations represented by SMIRKS
 - [c:10][C][N]>>[c:10][Cl]
- Nightly updates add new data on a continual basis

1. Warner *et al.* *J. Chem. Inf. Model.* **2010**, *50*, 1350-1357.

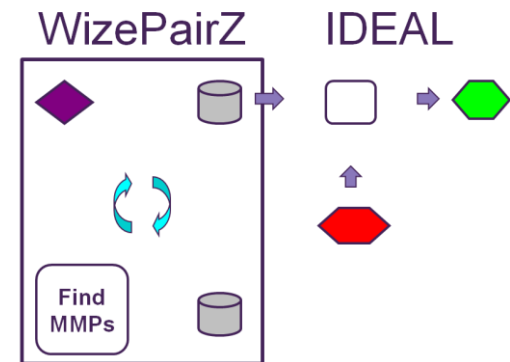
Chemical Context

- Transformations stored with varied degrees of local context

- 1 bond – Most general
- 2 bond
- 3 bond
- 4 bond – Most specific



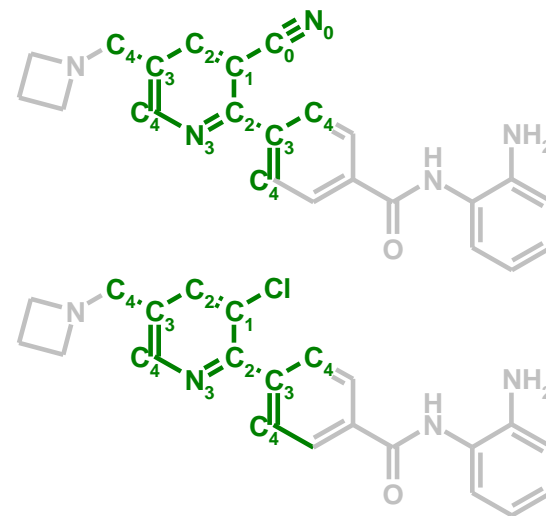
Chemical Context



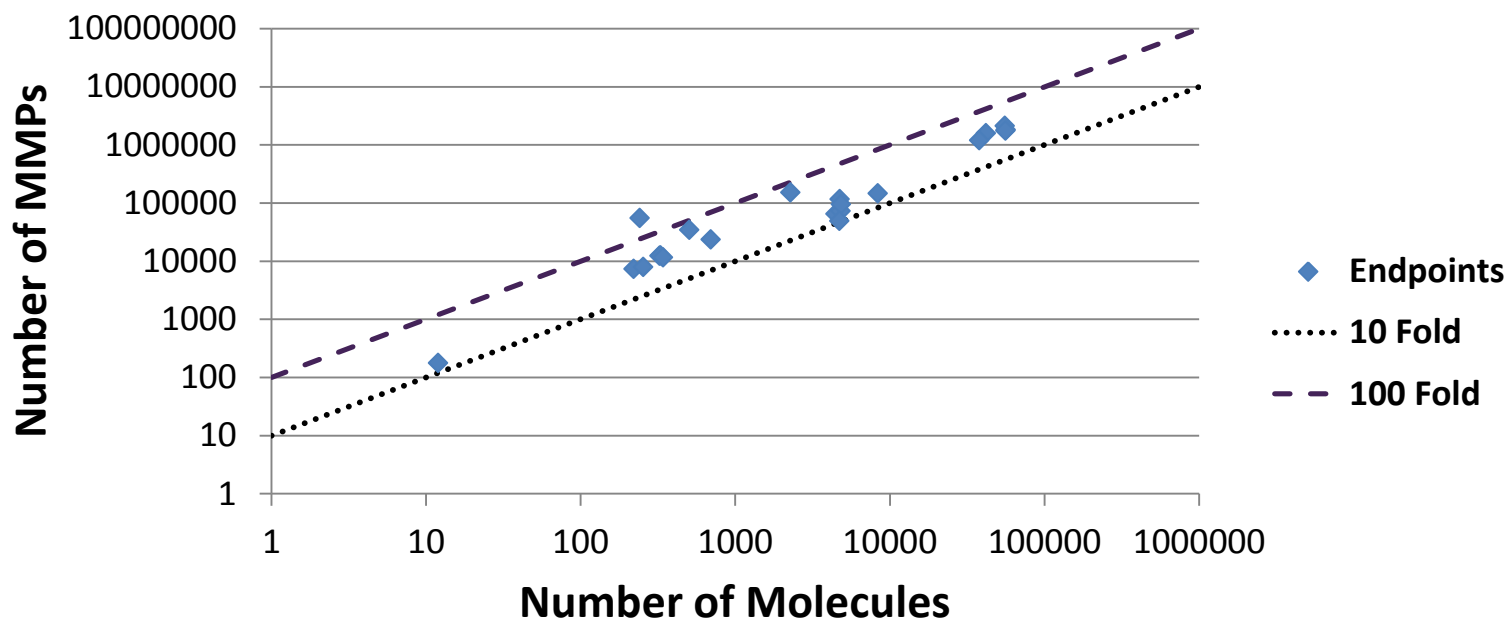
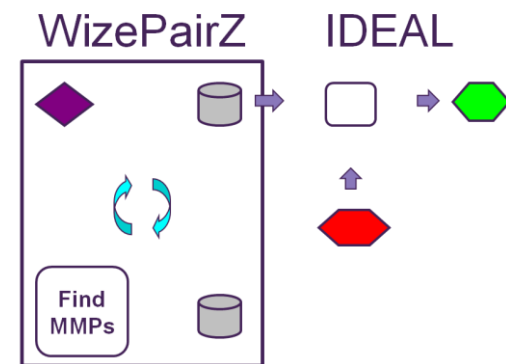
- Transformations stored with varied degrees of local context

- 1 bond – Most general
- 2 bond
- 3 bond
- 4 bond – Most specific

- Transformations and their contexts encoded as a single entity

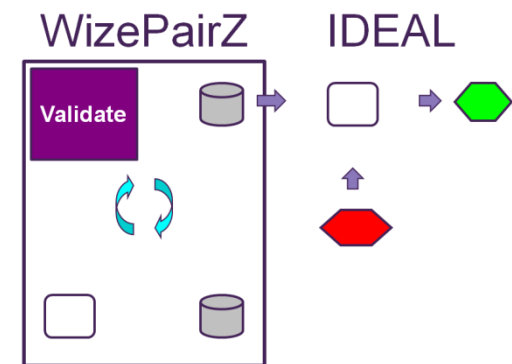


Many Matched Pairs



- Approximately linear relationship between N compounds and N MMPs
 - Between 10 fold and 100 fold as many MMPs as molecules
- Approximately 2 MMP observations per transformation

Transformation Quality Metrics



- **Quantitative approaches**

- Means and standard deviations
- 95% confidence intervals
- Not applicable to out-of-range measurements

- **Probabilistic approaches**

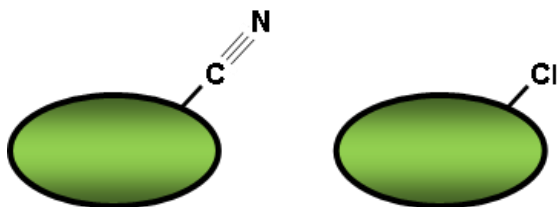
- Estimate the likelihood that a transformation will do something useful (or harmful) when applied to a new compound

- **Comparison to Known Distributions**

- Hajduk *et al.* ¹

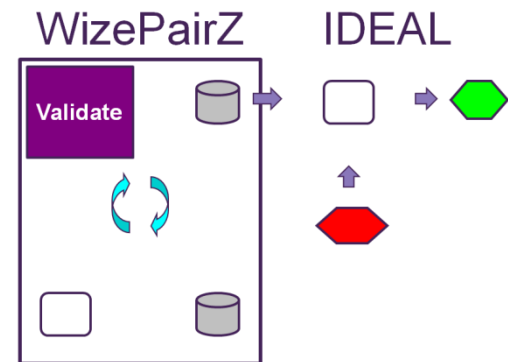
1. Hajduk *et al.* *J. Med. Chem.* **2008**, *51*, p553.

Transformation Quality Metric

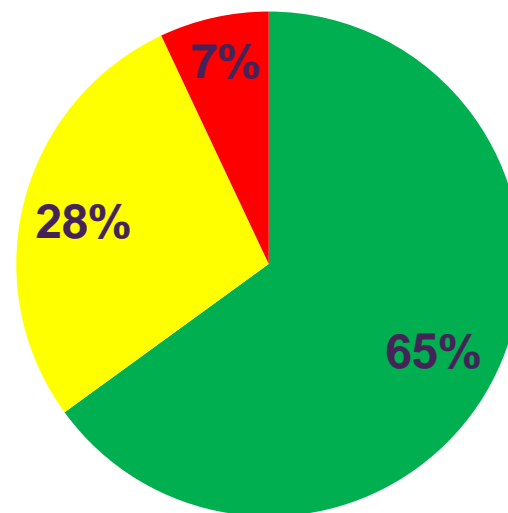


N observations

- **Probabilistic Approach:** Represents likelihood of success on applying transformation to a new compound
- Success defined as...
 - A change of greater than experimental error
 - Typically assumed to be 0.3 log units
- Expressed as a percentage
- Can handle out-of-range measurements
 - 5.0 to <4.5 – successful transformation
 - <5.0 to <4.5 – not a successful transformation

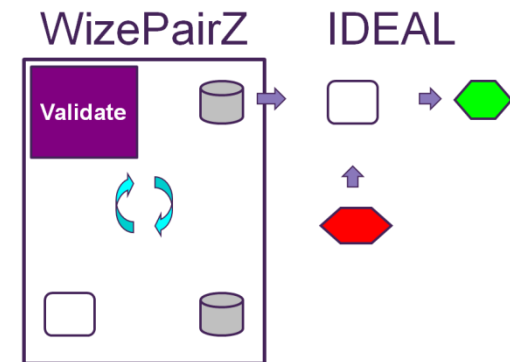


Transformation Profile



- Improve Property
- No Change
- Worsen Property

Cumulative Density Function (CDF)



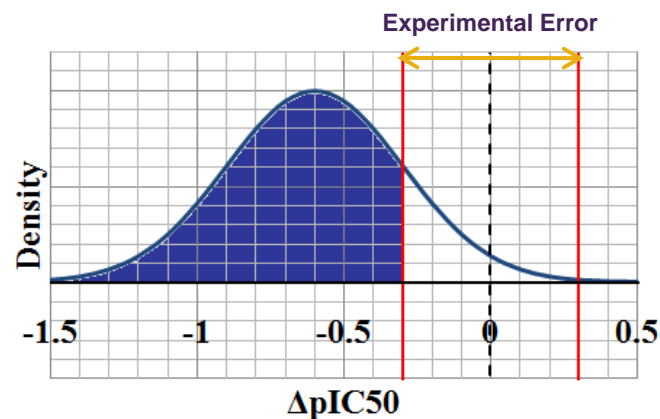
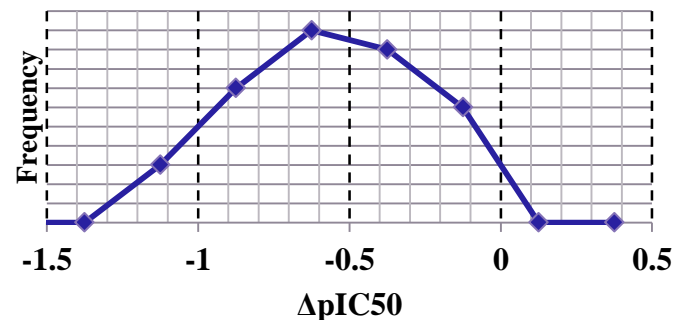
- Assume normal distribution for a set of ΔpIC_{50} values
 - Mean and Std_Dev from Observations

- Estimate probability of success with the CDF

$$CDF(x, \mu, \sigma) = \frac{1}{2} \left[1 + erf \left(\frac{x - \mu}{\sigma \sqrt{2}} \right) \right]$$

where *erf* is the error function, expressed as a Taylor expansion

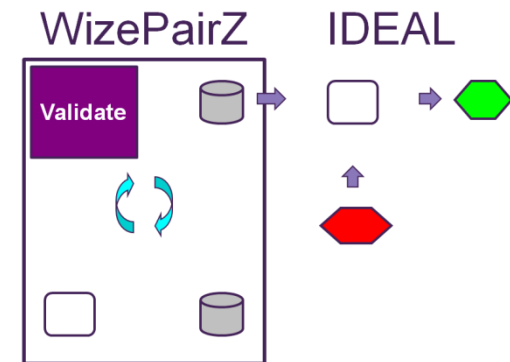
Distribution of 35 ΔpIC_{50} Observations for a Specific Transformation



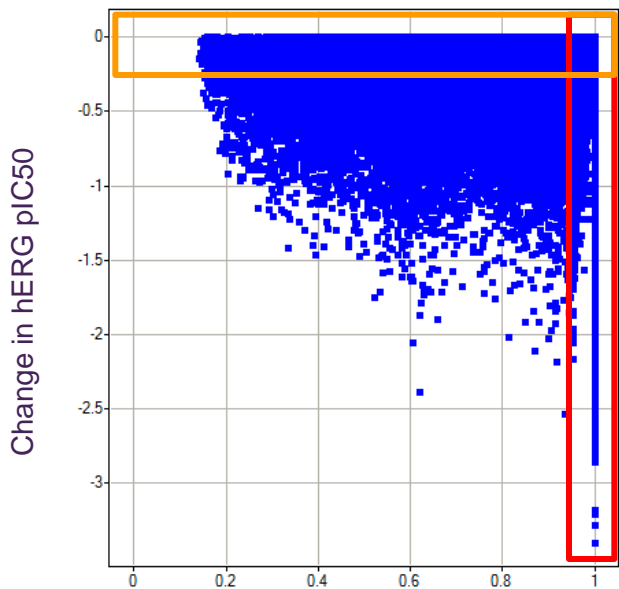
CDF = 0.84

Effect of N and Sim for hERG Data

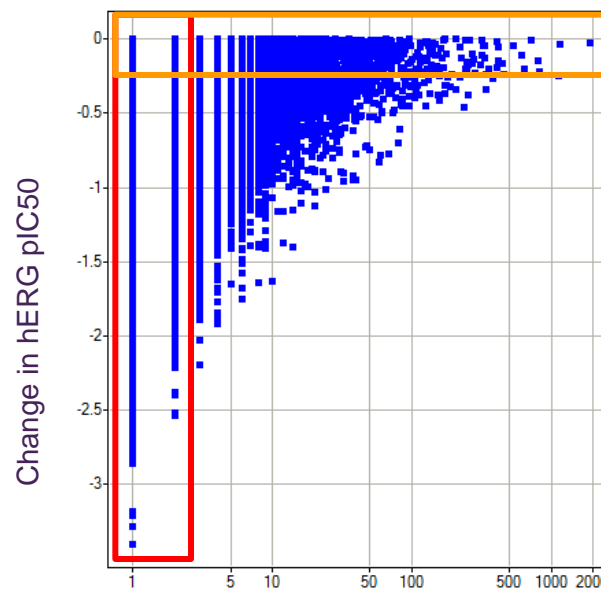
Sim is the average pairwise LINGO¹-Tanimoto similarity between all cores in the set of observations



Small changes



Sim (Lingo Tanimoto)



N (Number of Observations)

Low confidence

1. Vidal *et al.* (2008) *JCIM*, 45, p386

Corrected CDF (CCDF)

$$CCDF = CDF \times N_{adjust} \times SIM_{adjust}$$

$$N_{adjust} = 1 - e^{-aN}$$

- **Sample Size Adjustment:**

- Demote transformations that have only limited MMP observations

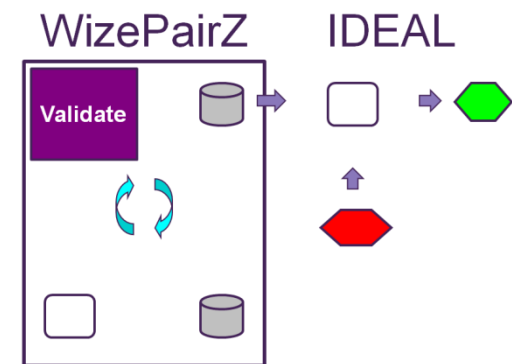
$$SIM_{adjust} = e^{-bSim}$$

- **Similarity Adjustment:**

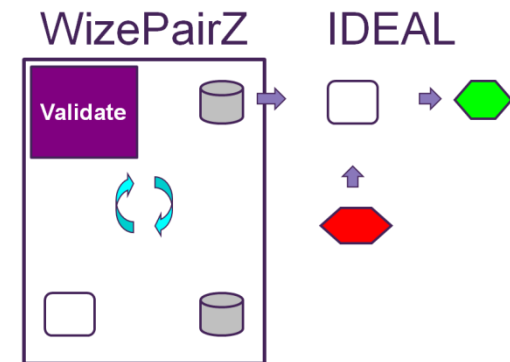
- Demote transformations seen only in homogenous chemistries
- *Sim* is the mean pairwise similarity: (LINGOS with Tanimoto)

- Functions chosen by eyeballing the data

- Parameterization done with an internal, re-sampling validation experiment



Parameterizing the CCDF

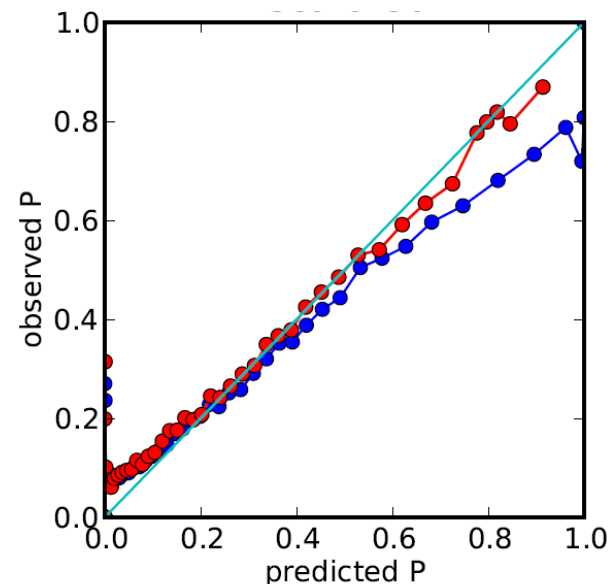


- hERG transformations:
 - Optimal $a = 1.0$; Optimal $b = 0.2$

$$CCDF(\text{paramd}) = CDF \times e^{-N} \times e^{-0.2Tan}$$

- CCDF probability estimates are more accurate than the uncorrected CDF

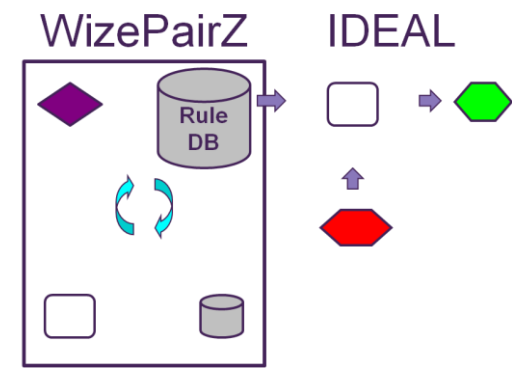
Expected versus observed probabilities of success, determined from an external test set



CDF ● CCDF



hERG Transformations



- Inspired by International Panel on Climate Change Report ¹
- Number of hERG transformation rules that satisfy each condition

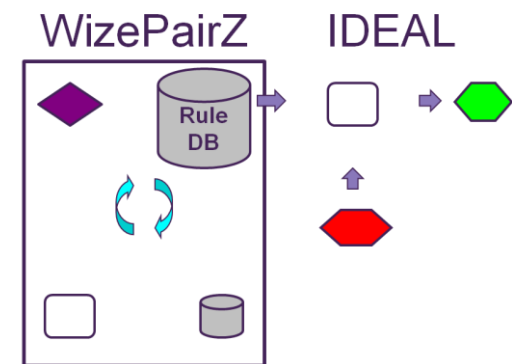
Condition (CCDF)	Label	Number of Rules
Greater than 0.85	Very Likely	979
0.65-0.85	Likely	32706
0.50-0.65	More likely than not	51111
0.33-0.50	Possible_1_in_3	112293
Less than 0.33	Unlikely	649751

- Many fewer rules than would be found with 95% confidence intervals

1. International Panel on Climate Change, Fourth Assessment Report **2007**

hERG Transformations

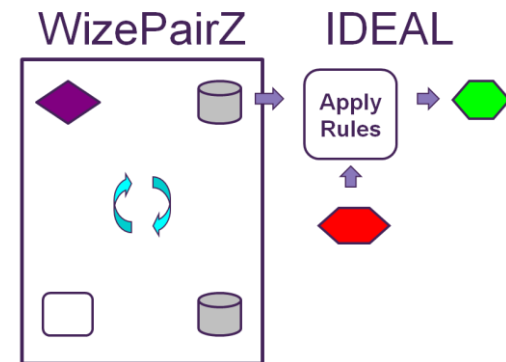
- Some top hERG lowering transformations



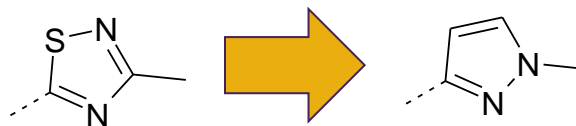
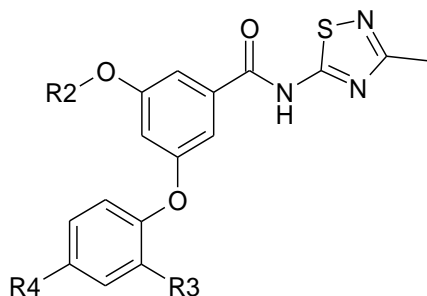
Before	After	Mean Δ	StdDev Δ	N	SIM	CDF	CCDF
	<chem>C-NH2</chem>	-0.58	0.08	6	0.29	1.00	0.97
		-0.76	0.12	7	0.40	1.00	0.96
		-0.85	0.19	15	0.43	1.00	0.96
		-0.64	0.13	15	0.39	1.00	0.96
		-0.84	0.21	12	0.40	1.00	0.96
		-0.83	0.20	5	0.34	1.00	0.96

Glucokinase Activators

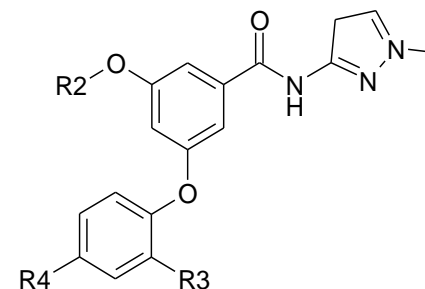
Retrospective Project Example



- Query submit through IDEAL (a virtual design platform):
 - Request: Fix my hERG problem whilst maintaining my potency



- Very likely to reduce hERG (90%)
- Unlikely to reduce pEC₅₀ (20%)



Δ pEC ₅₀	=	0.0
Δ logD	=	-0.6
Δ hERG pIC ₅₀	=	-0.6

- Waring, M. J. (2010) Optimisation of Neutral Glucokinase Activators. The Discovery of AZD1092 Presented at the Gordon Research Conference on Medicinal Chemistry, August 8-13, 2010.
- Johnstone, C. et al. (2010) The Medicinal Chemistry of Glucokinase Activators: Property-based Drug Discovery. Presented at EFMC-ISMIC 2010 – XX1st Int. Symposium On Medicinal Chemistry, Brussels, Belgium, September 5-9, 2010.

In Conclusion

- Pharmaceutical databases contain a wealth of information
 - MMPA can be used to exploit this information
- WizePairZ uses MMPA to mine our databases for interesting molecular transformation 'rules'
- Molecular transformations in WizePairZ are validated with the Corrected Cumulative Density Function
 - Parameterized to promote transformations proven many times over diverse chemistries

Acknowledgements

■ Contributors

- Ed Griffin
- Steve St-Gallay
- Dan Warner

■ WizePairZ Team

- Craig Bruce
- Martin Harrison
- Chris Green
- Attila Ting

■ Input

- Andrew Leach
- Al Rabow
- Andy Davis
- Sarah Aaron

■ Thank You!

