

WizePairZ

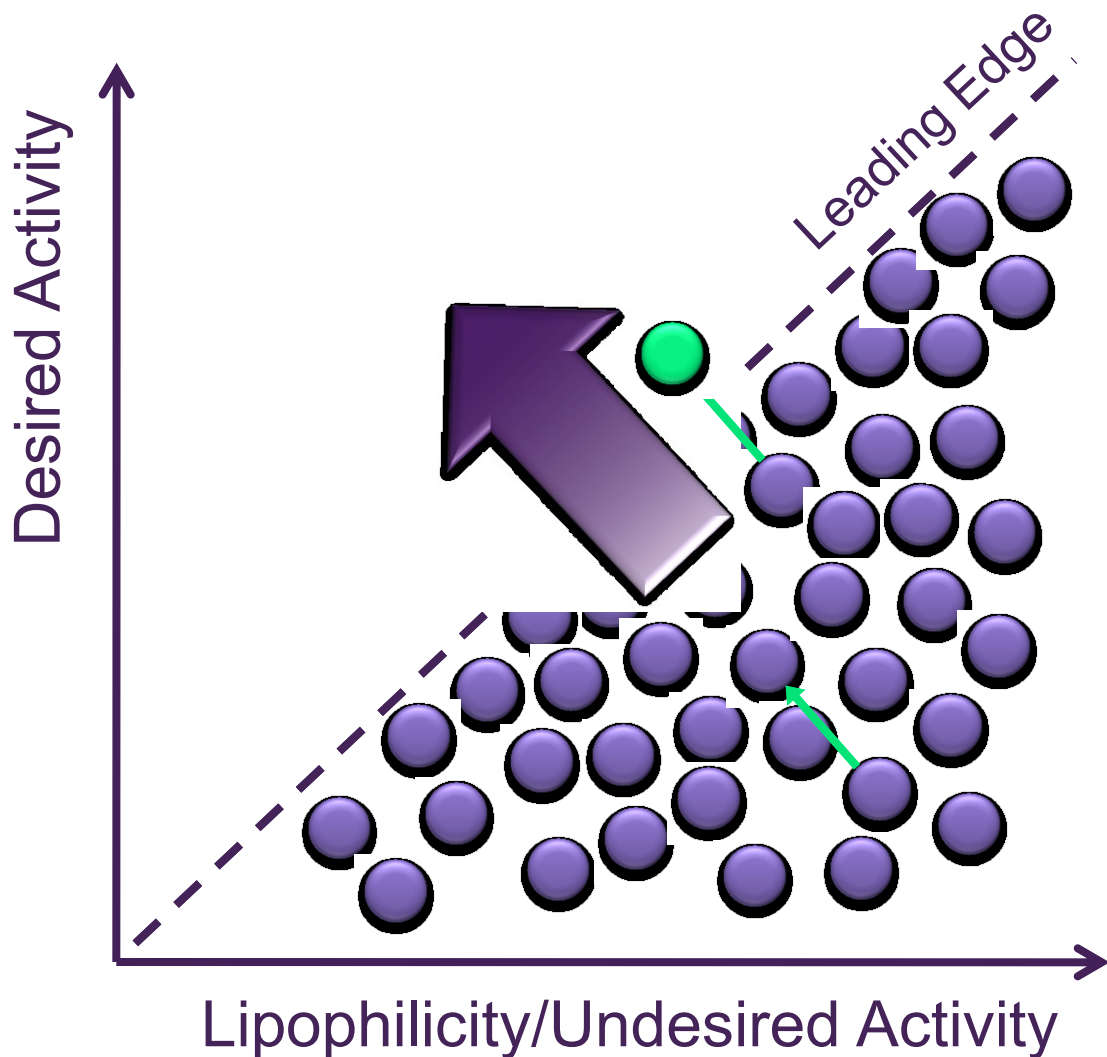
A Novel Algorithm to Identify, Encode and Exploit Matched Molecular Pairs with Unspecified Cores in Medicinal Chemistry

Steve St-Gallay, Dan Warner, Ed Griffen, David Wood

Introduction

- Overview of the WizePairZ approach
 - Pushing the 'leading edge'
 - Assuming additivity
- WizePairZ applied to a public dataset
- Implementation of WizePairZ
- Selecting significant transformations

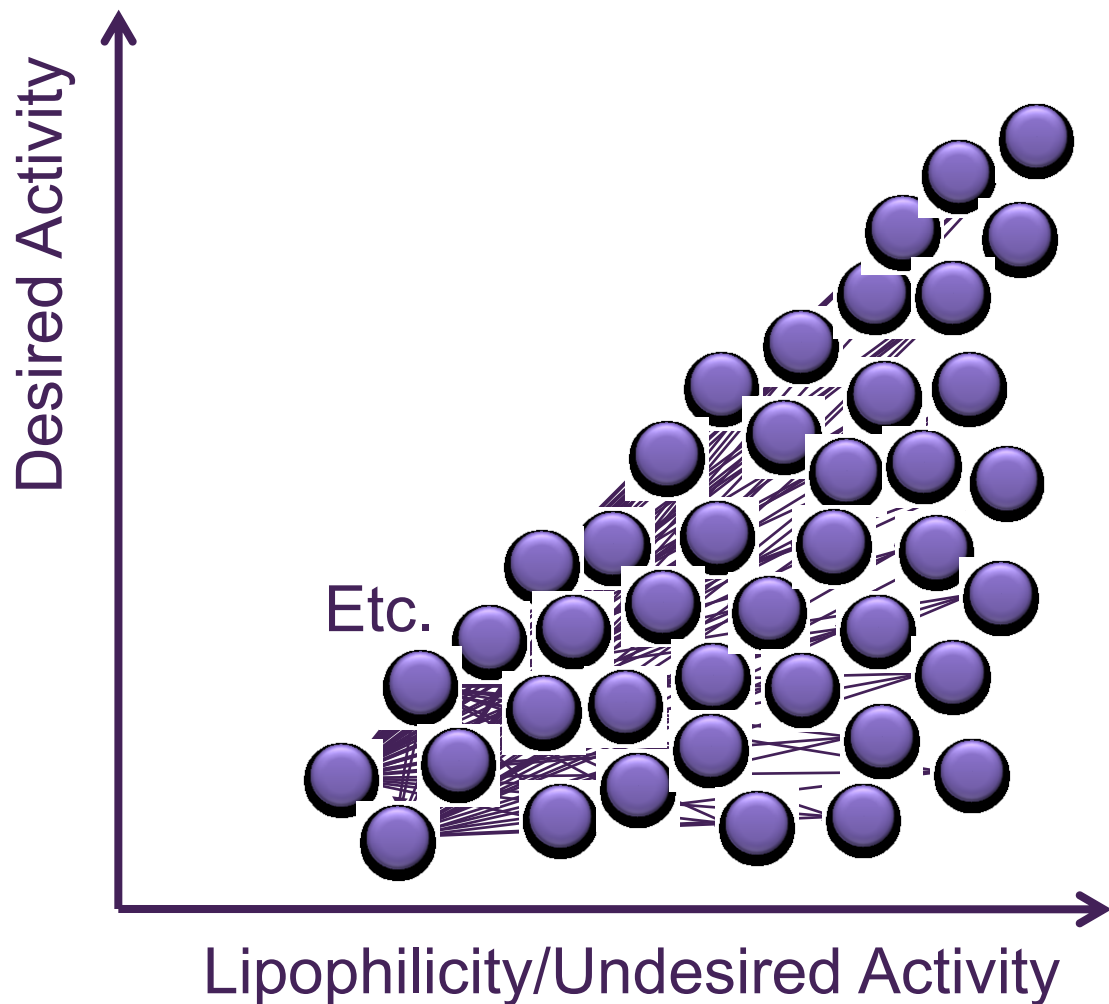
Problem with Drug Discovery Projects



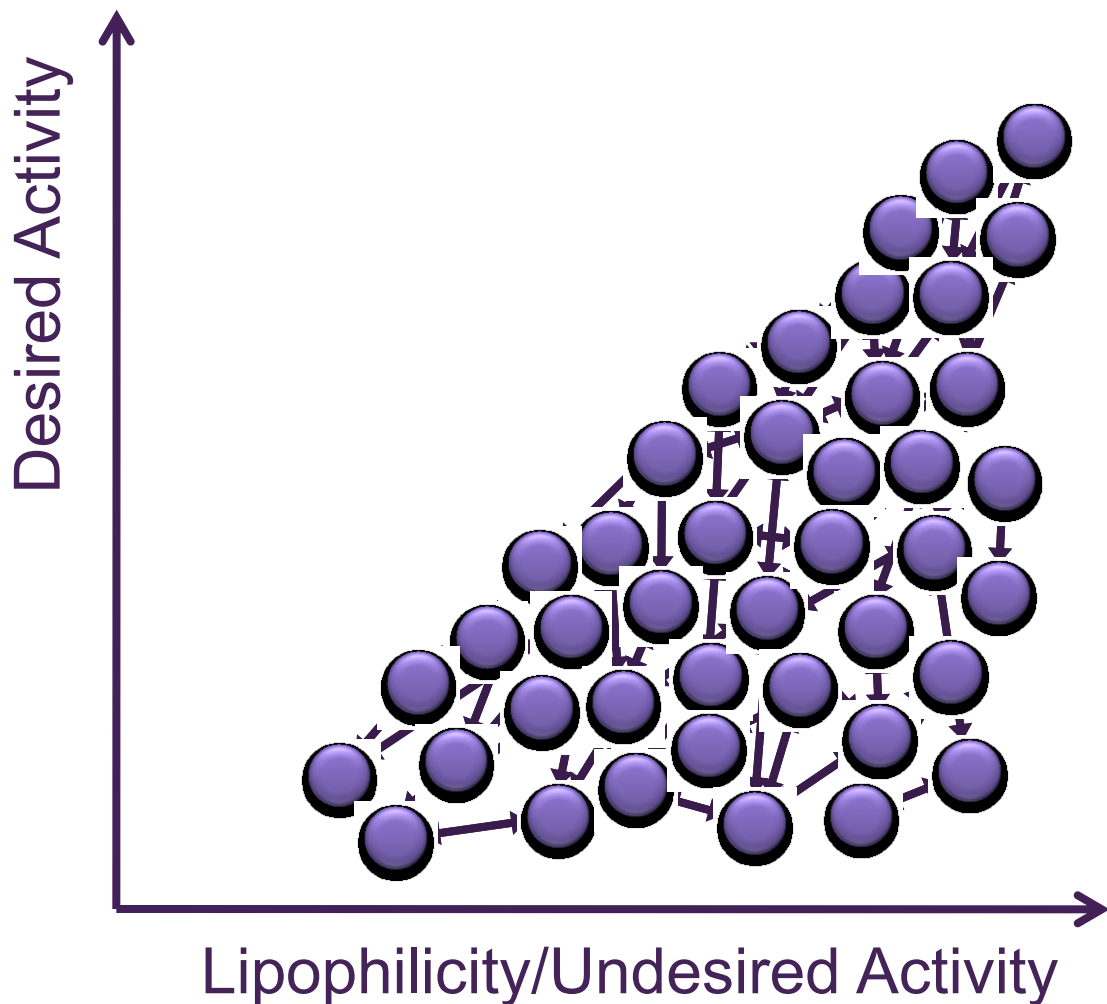
- Desired activities (such as potency) can correlate with undesired properties (such as hERG), through lipophilicity
- Achieving a breakthrough requires pushing the 'leading edge'
- Properties of molecules are also governed by molecular structure, not just lipophilicity
 - Transformations orthogonal to the leading edge
- Exploit the structural variations to push the leading edge

WizePairZ - Overview

- Performs match pairs for each compound in the set

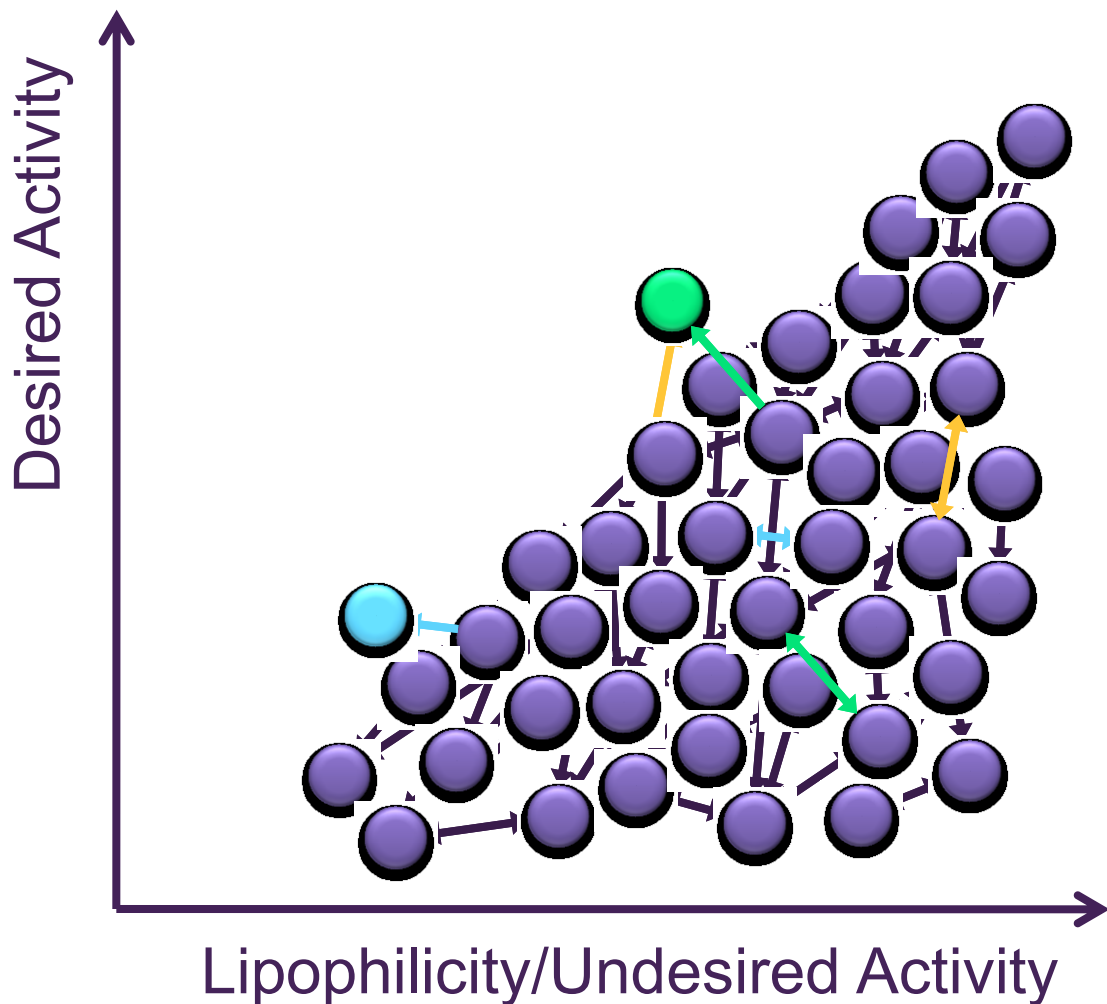


WizePairZ - Overview



- Performs match pairs for each compound in the set
- Looks at the maximum common substructure for each pair and throws the pair away if there is less than 90% in common
- Store the transformation and associated change in properties

WizePairZ - Overview



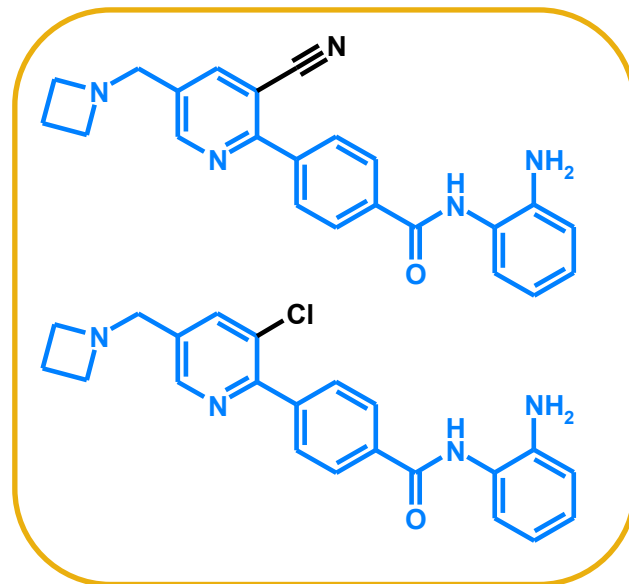
- Performs match pairs for each compound in the set
- Looks at the maximum common substructure for each pair and throws the pair away if there is less than 90% in common
- Store the transformation and associated change in properties
- Apply the transformations to the existing compounds
 - Can any existing compounds be modified (pair does not exist?)
 - Do they push the leading edge?
- Multiple transforms suggesting the same compound

Histone Deacetylase Inhibitor data set*

Structure	R1	R2	Compound	logD _{7.4}	HDAC pIC ₅₀
		CN	3	1.58	7.77 ± 0.69
		CH ₃	13a	1.18	7.67 ± 0.14
		CN	13b	1.22	8.01 ± 0.35
		Cl	13c	NV	7.73 ± 0.06
		F	13d	1.69	7.30 ± 0.24
		CH ₃	14a	1.20	7.73 ± 0.40
		CN	14b	1.14	8.01 ± 0.18
		Cl	14c	1.81	7.79 ± 0.25
		F	14d	NV	7.42 ± 0.35
		CH ₃	15a	1.44	7.72 ± 0.16
		CN	15b	1.28	7.98 ± 0.07
		Cl	15c	1.68	7.88 ± 0.79
		F	15d	1.60	7.48 ± 0.08

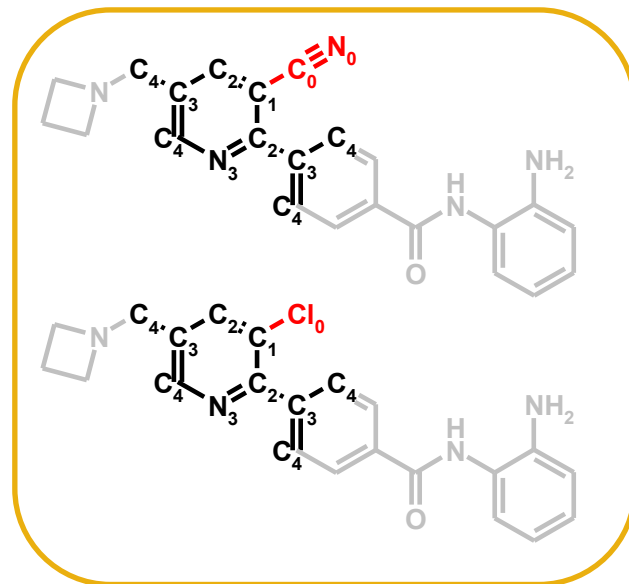
Identification of Potential Matched Pairs

- Do the MCS



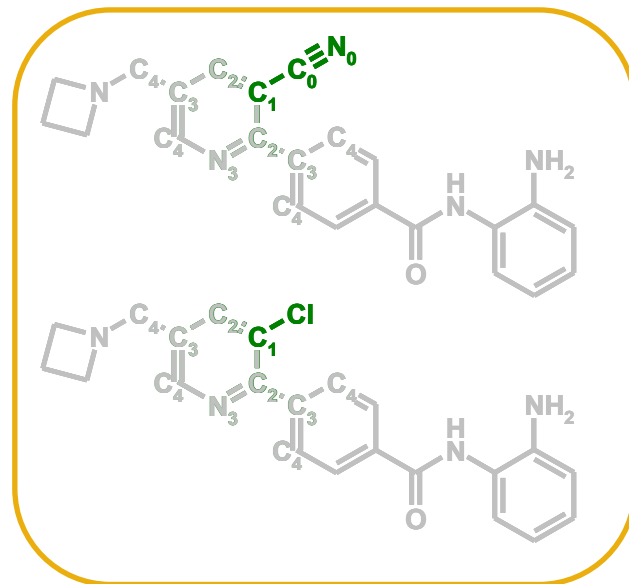
Match Pair Verification

- Do the MCS
- Spot the difference



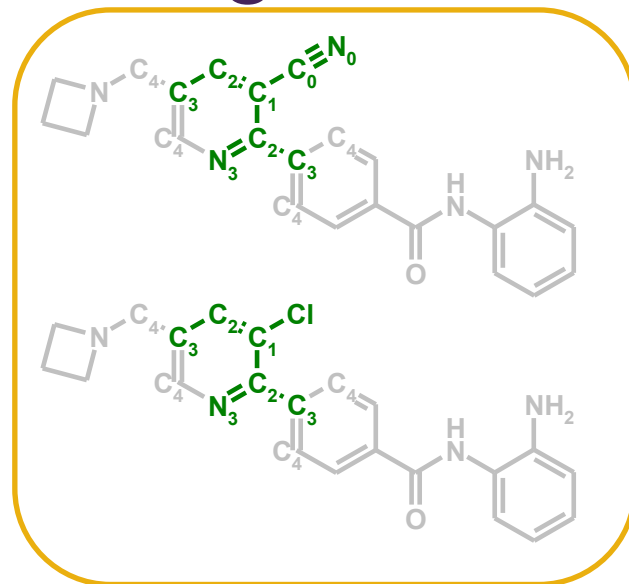
Definition of Local Environment Shells

- Do the MCS
- Spot the difference
- Levels of chemical environment for the transformation
 - Up to 4 fragments for each difference



Encoding the SMIRKS String

- Do the MCS
- Spot the difference
- Levels of chemical environment for the transformation
 - Up to 4 fragments for each difference
- Generate the SMIRKS



[c:6][c:11]([H])[c:10]([c:9]([n:8])[c:12])[Cl]>>[c:6][c:11]([H])[c:10]([c:9]([n:8])[c:12])[C]#[N]

Aggregation of Transformations

- Do the MCS
- Spot the difference
- Levels of chemical environment for the transformation
 - Up to 4 fragments for each difference
- Generate the SMIRKS
- Remove atom mappings from SMIRKS

[c:7]([H])>>[c:7][F]

[c:18]([H])>>[c:18][F]

cH.cF

Aggregation of Transformations

- Do the MCS
- Spot the difference
- Levels of chemical environment for the transformation
 - Up to 4 fragments for each difference
- Generate the SMIRKS
- Remove atom mappings from SMIRKS
- Combine all identical transformations
 - Associate with data for the transformation

Number of observations (N)

The number of compound pairs that generate a transformation

Lingo Tanimoto (Sim)

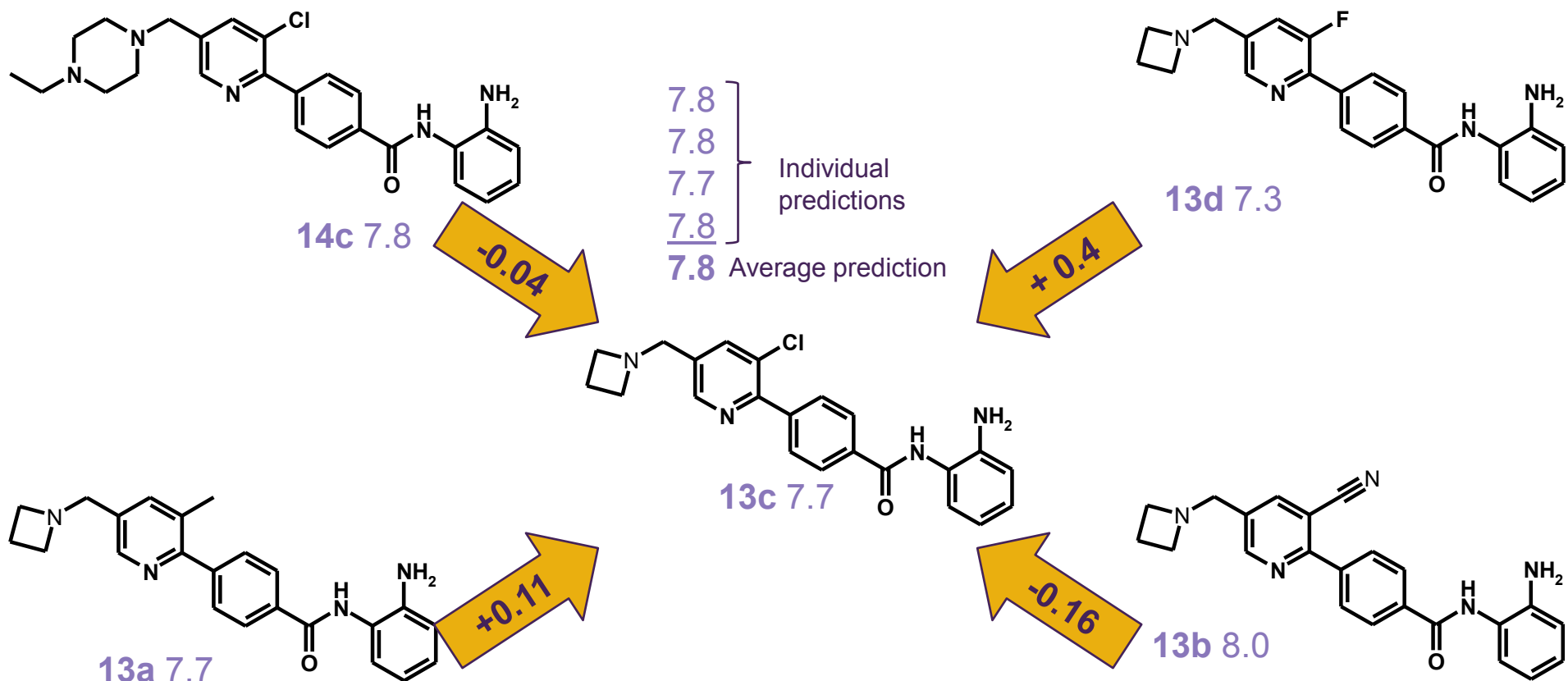
The average Tanimoto using Lingos between all the compounds in the set that generates a single transformation

Application of Extracted Transformations

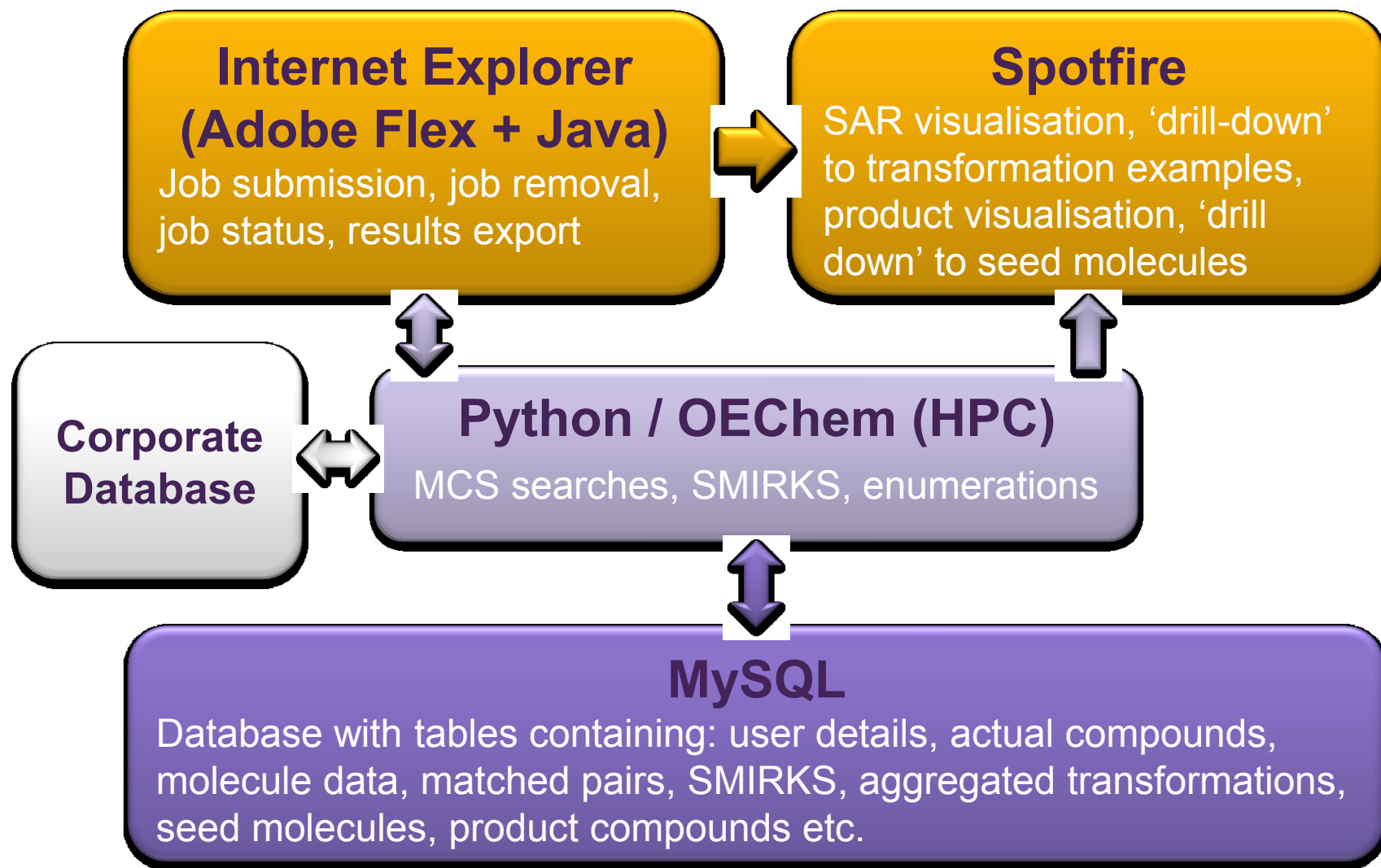
- Do the MCS
- Spot the difference
- Levels of chemical environment for the transformation
 - Up to 4 fragments for each difference
- Generate the SMIRKS
- Remove atom mappings from SMIRKS
- Combine all identical transformations
 - Associate with data for the transformation
- Apply transformations on seed compounds to enumerate virtual libraries
 - Estimate property values

Results

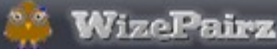
- 18 unique transformations generated for HDAC data set
- Application to seed compounds identified compound 13c (and 14d – not shown) multiple times (not used to generate transformations)



WizePairZ Platform



WizePairZ Interface


Welcome Steve

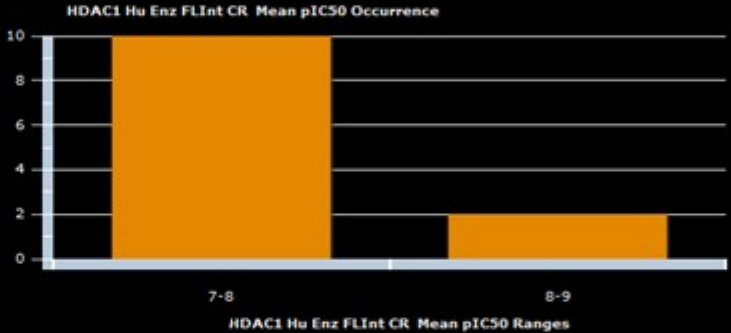
[Ibis Data](#) | [Generate](#) | [Tasks](#) | [Published Data](#) | [Admin](#) | [Help](#) | [About](#)

IBIS Information

Enter an IBIS key for your data:

Target Properties

- Paper Reference
- Molecule Formula Weight
- HDAC1 Hu Enz FLInt CR Mean pIC50
- Hu HCT116 Prolif Ttz CR Mean pIC50
- Molecule ACD LogD (7.4)
- hERG Hu CHO IF EPhs CR Mean pIC50
- Solubility Solid pH=7.4 Mean pSolubility
- Lipophilicity Octanol pH=7.4 Mean LogD

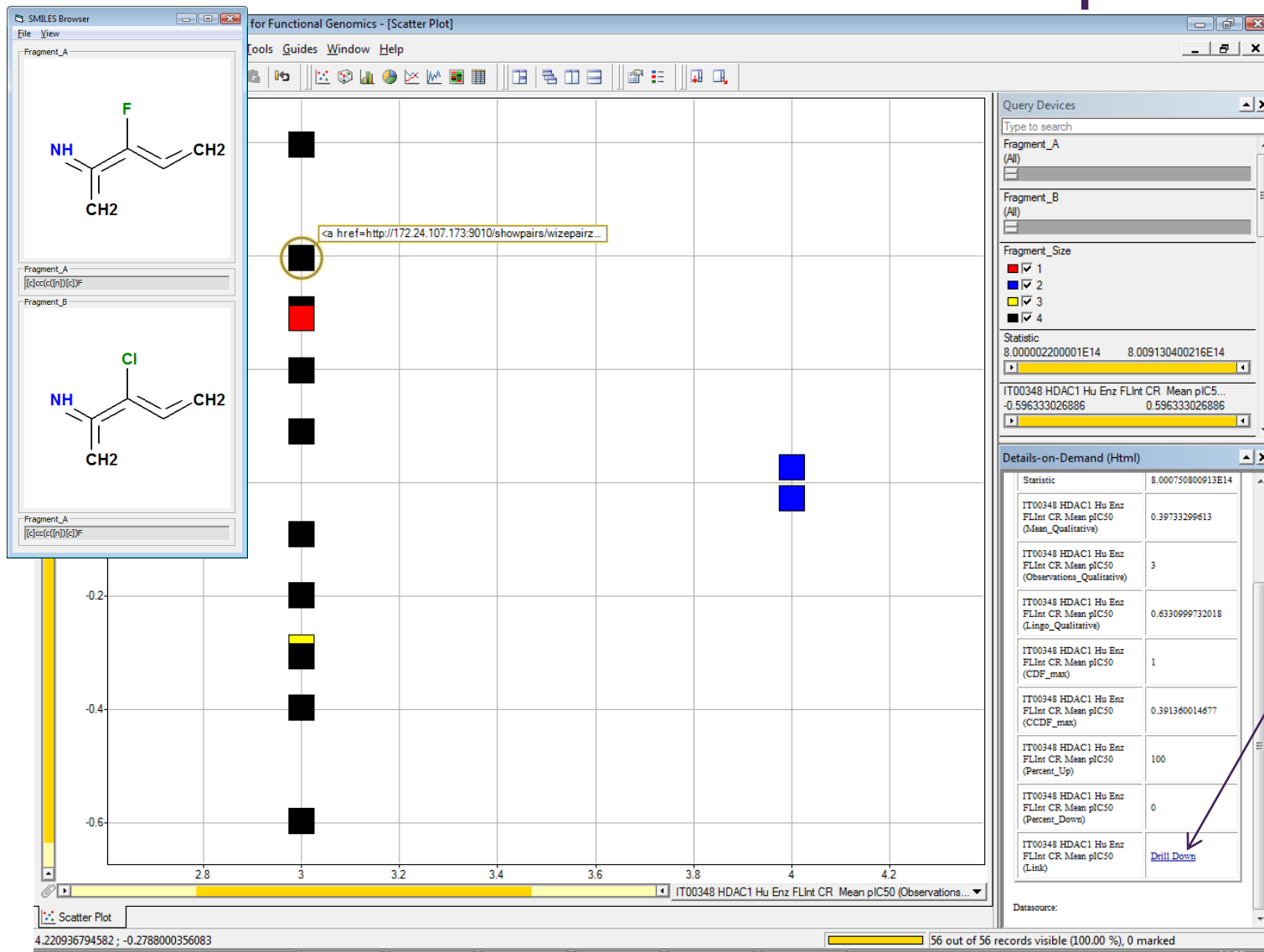


HDAC1 Hu Enz FLInt CR Mean pIC50 Ranges

IBIS Summary Information 12 Compounds

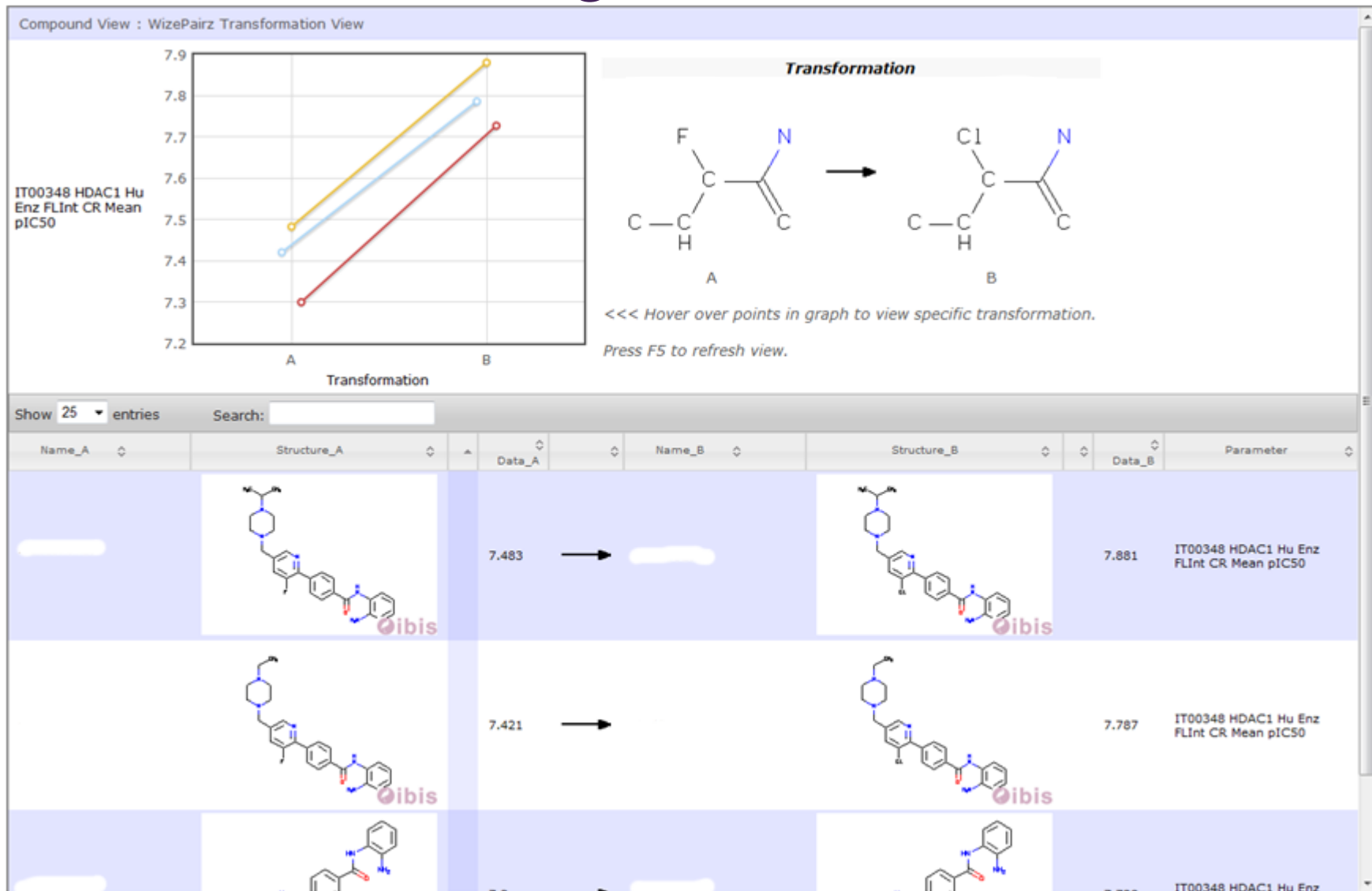
Compound	Compound	Paper Ref	Molecule	Qualifier	Qualifier	Molecule	Qualifier	Qualifier	Qualifier	Qualifier	Qualifier
	CC(C)N1c	15b	454.58	7.976	6.845	1.42	4.715	2.859		1.28	
	CCN1CC	14b	440.55	8.012	6.858	1.09	<	4.672	<	2.712	1.14
	c1ccc(c(c)1	13b	383.45	8.005	6.857	0.38	<	4.562		3.127	1.22
	CCN1CC	14c	449.98	7.787	6.469	2.27	5.084	NV			1.81
	CC(C)N1c	15c	464.01	7.881	6.511	2.60	5.148	<	2.827		1.68
	c1ccc(c(c)1	13c	392.89	7.728	6.524	1.50	4.652		2.873	NV	
	Cc1cc(cnc1	15a	443.59	7.72	6.623	2.31	<	4.544	<	2.582	1.44
	CCN1CC	14a	429.57	7.733	6.638	1.98	<	4.515	<	2.629	1.2

View WisePairZ Results in Spotfire

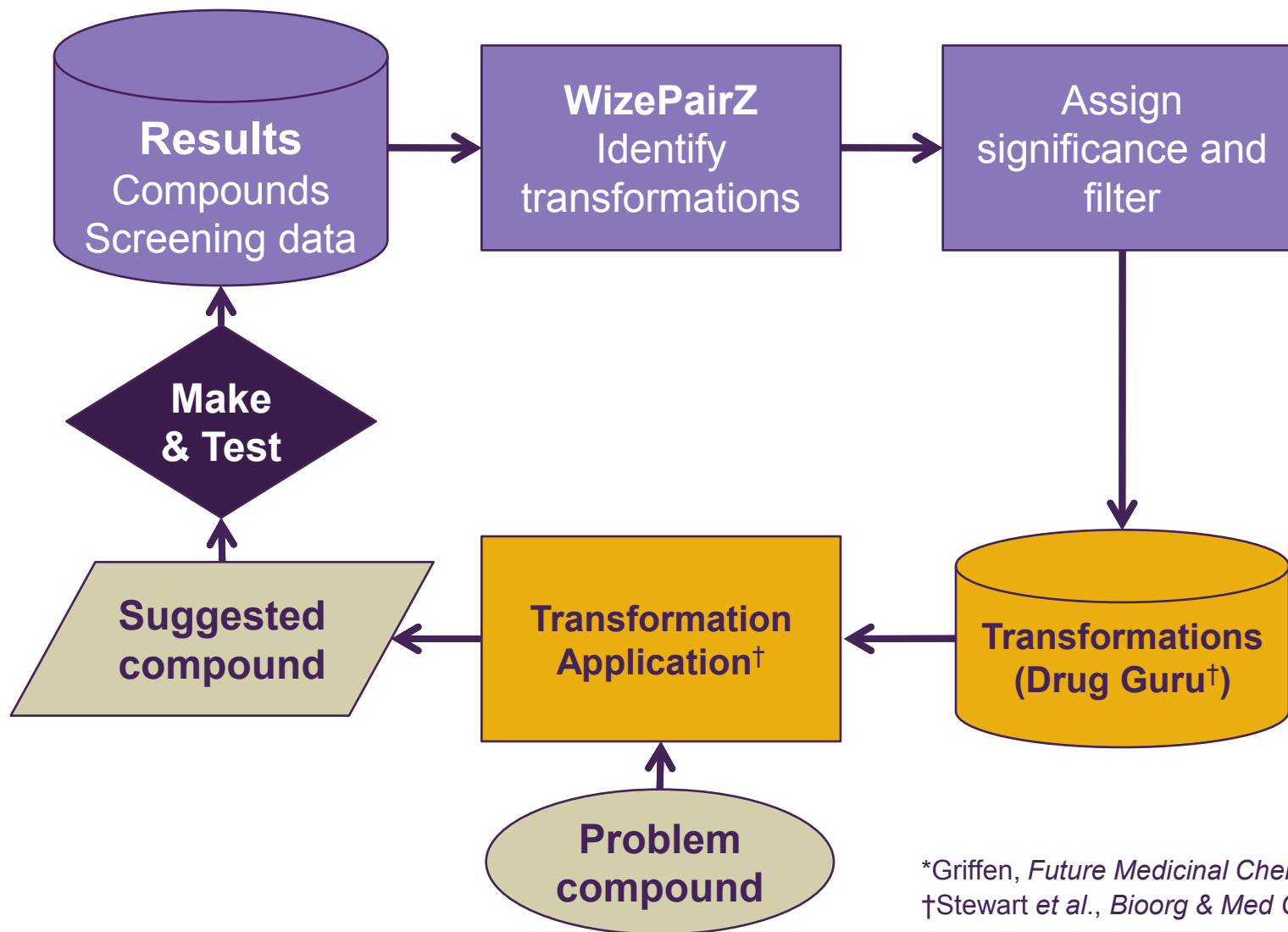


Select a transformation and drill-down to the original data with this hyperlink

Drill-down to Original Pairs Data



Incorporation*



*Griffen, *Future Medicinal Chemistry* 1 (2009) 405
†Stewart et al., *Bioorg & Med Chem* 14 (2006) 7011

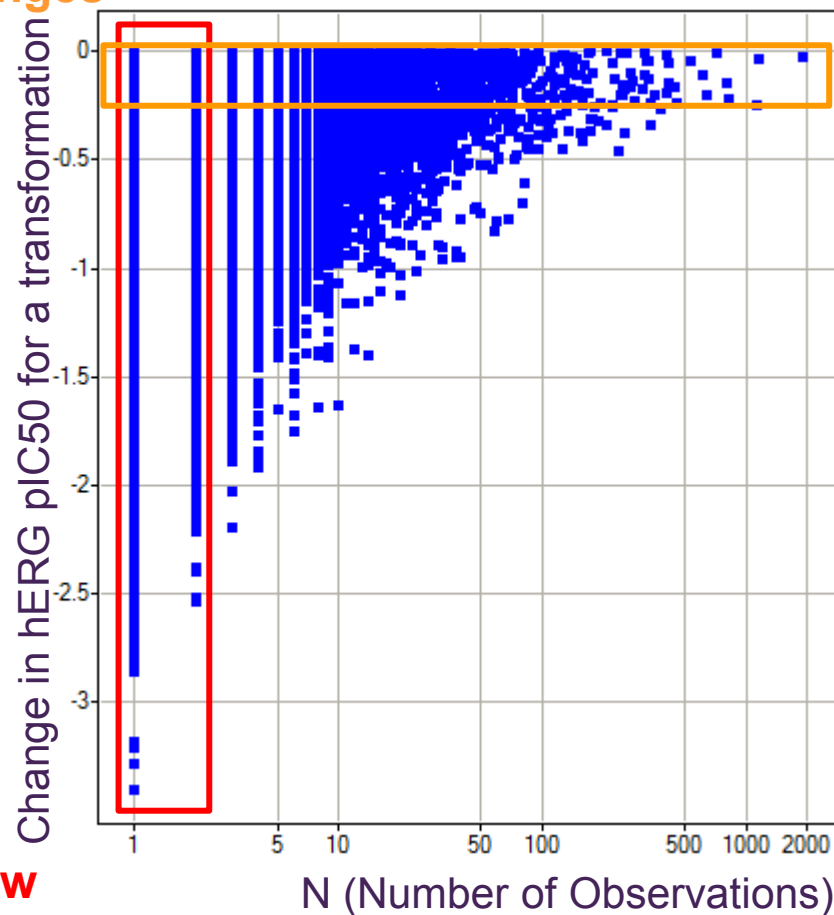
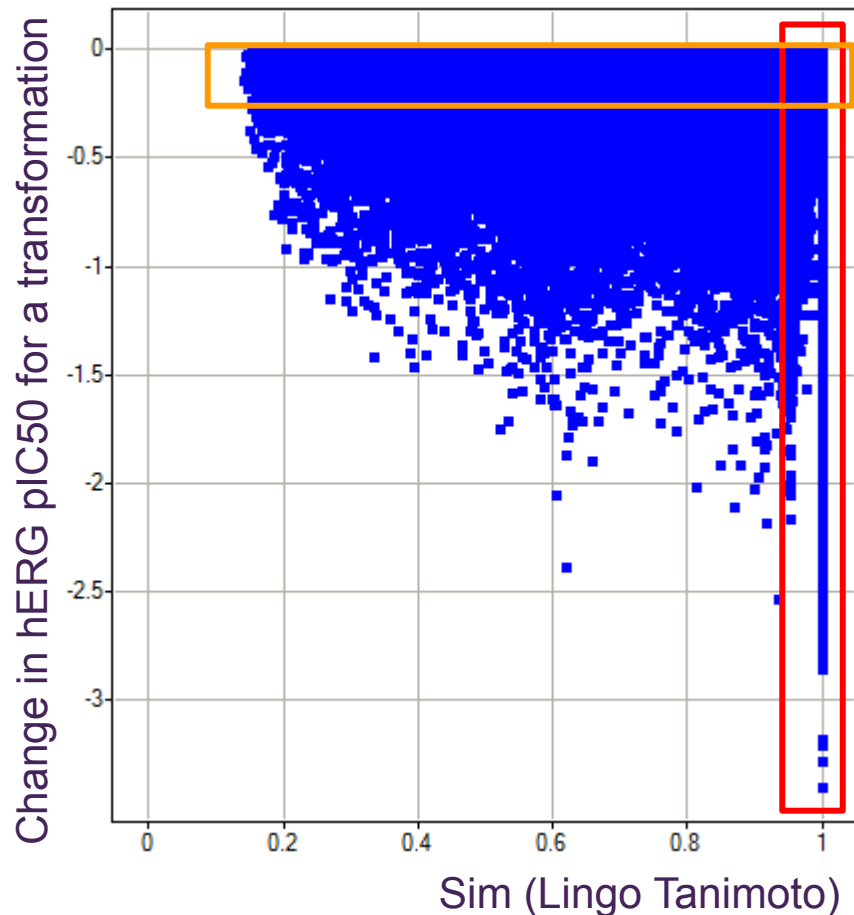
Assign Significance and Filter

- Select a desired cut-off for the property change
 - E.g. want the potency to change by >0.3 log units
- Estimate a probability that the transformation will result in the desired property change
 - Based on the Cumulative Distribution Function
 - Modified by
 - Number of times the transformation has been observed
 - The Lingo Tanimoto distance between the compounds producing the transformation
 - More observations and more diversity engenders more confidence
- Select the probability you'd accept and use these transformations

Effect of N and Sim for hERG Data

- >40,000 IC_{50} measurements, ~350,000 transformations

Small
changes



Low
confidence

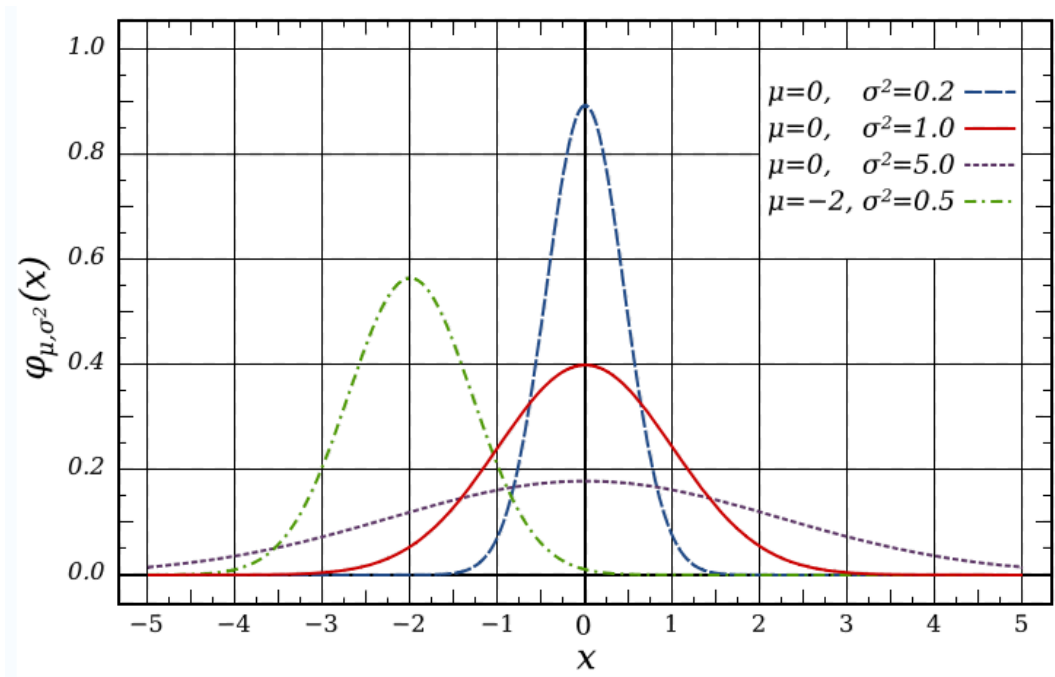
Cumulative Distribution Function (CDF)

- Based on an assumed normal distribution

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

Where erf is the error function, which can be expressed as a Taylor expansion

$$\begin{aligned} \operatorname{erf}(z) &= \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n+1}}{n!(2n+1)} \\ &= \frac{2}{\sqrt{\pi}} \left[z - \frac{1}{3} z^3 + \frac{1}{10} z^5 - \frac{1}{42} z^7 + \frac{1}{216} z^9 - \dots \right] \end{aligned}$$



Corrected Cumulative Distribution Function (CCDF)

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

Where

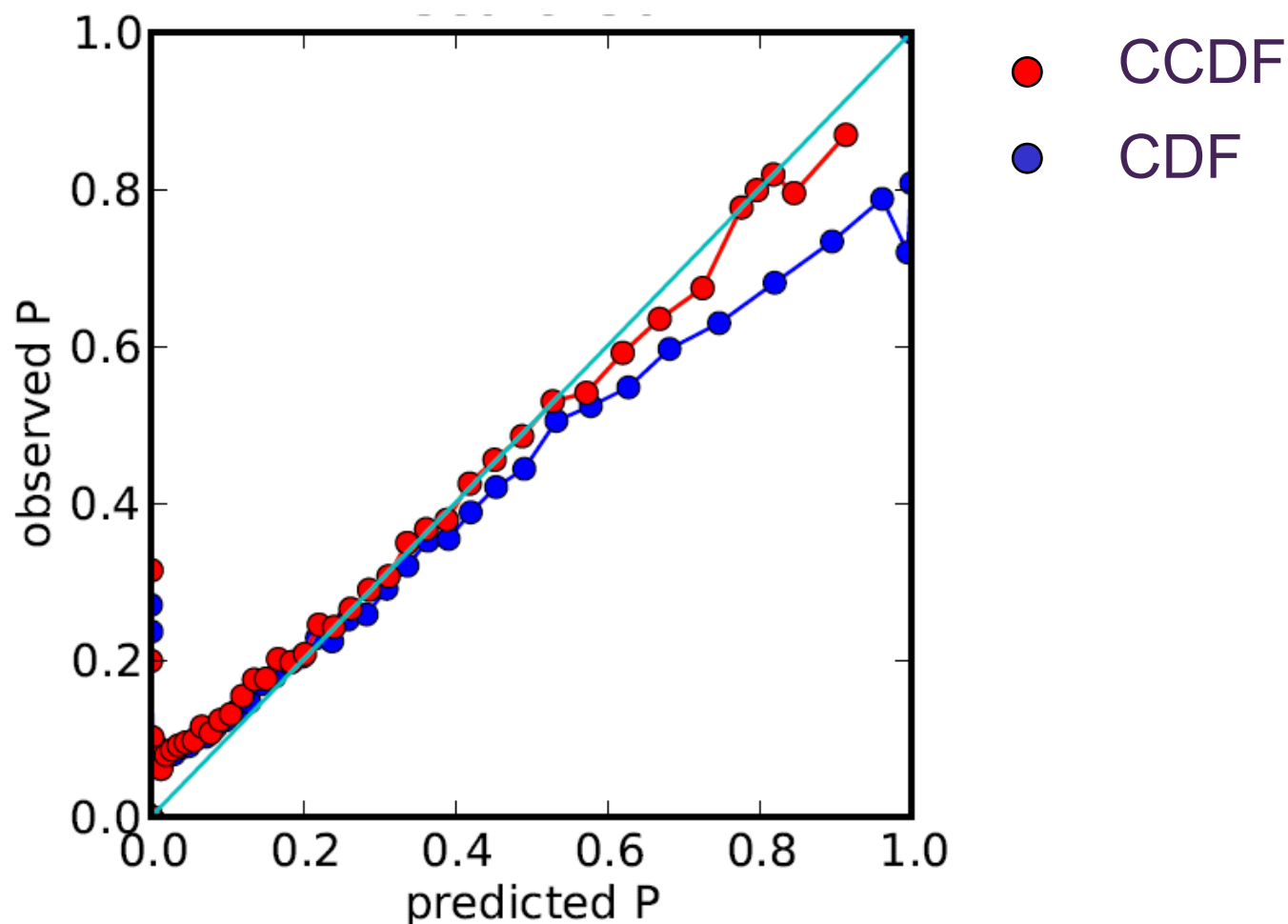
$$N_{adjust} = 1 - \exp^{-N}$$

$$Sim_{adjust} = \exp \frac{Sim}{10}$$

These functions were derived by eyeballing the data and optimising the coefficients with multiple data sets

- Comparison with an external test set demonstrated this function improves on the estimated probability that a given transformation would make the required change

Observed vs Predicted Probabilities for External hERG Test Dataset



Summary

- **How WizePairZ works**
 - Automatically identifies and extracts matched molecular pairs from a collection of compounds
 - Associates the transformations with property changes
 - Exploits the transformations to suggest compounds with improved properties
- **Application of WizePairZ to drug discovery**
 - Individual drug discovery project problems
 - Converting corporate databases into knowledge for DrugGuru-like systems
- **Assigning significance to transformations**
 - Adjust the cumulative distribution function with a term for the number of observations and another for the similarity

Acknowledgements

- Collaborators
 - Dan Warner
 - Ed Griffen
 - Dave Wood
- Other Input
 - Lilian Alcaraz
 - Craig Bruce
 - Chris Green
 - Austen Pimm
 - Barry Teobald
 - Brian Springthorpe
 - Attila Ting
 - Nicholas Tomkinson
 - Hitesh Sanganee
- Manuscript Proof Readers
 - Dave Cosgrove
 - Pete Kenny

SPARE

Cumulative Distribution 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

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

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

$$\begin{aligned} \operatorname{erf}(z) &= \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n+1}}{n!(2n+1)} \\ &= \frac{2}{\sqrt{\pi}} \left[z - \frac{1}{3} z^3 + \frac{1}{10} z^5 - \frac{1}{42} z^7 + \frac{1}{216} z^9 - \dots \right] \end{aligned}$$

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

