Data Visualisation: New Directions or Just Familiar Routes?

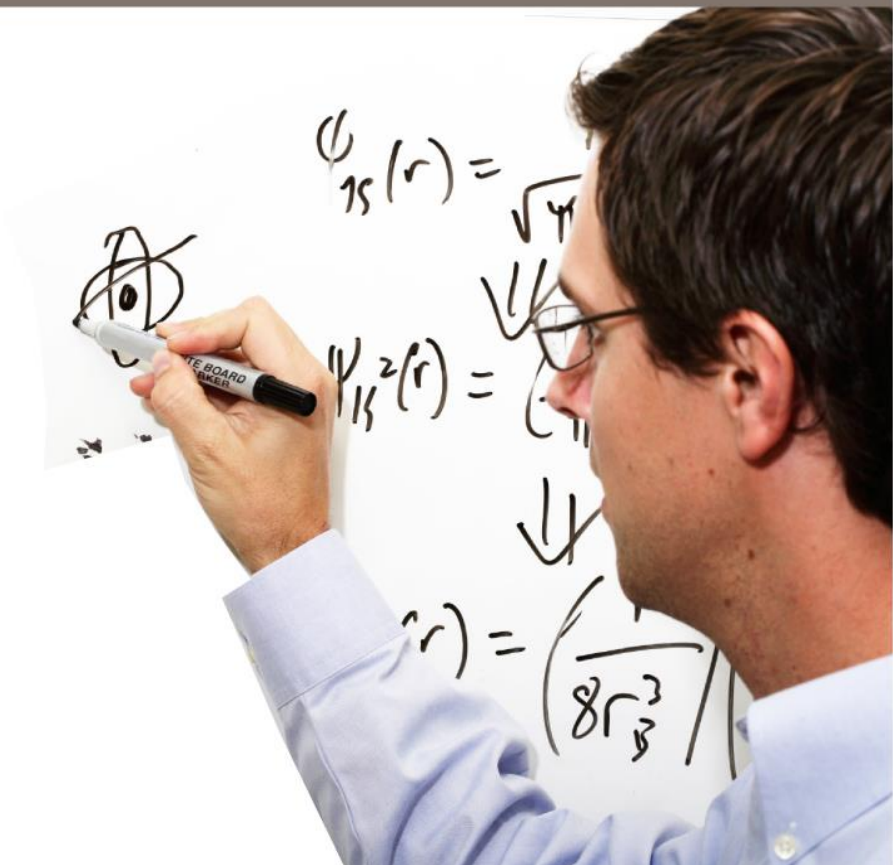
Ed Champness, Matt Segall & Peter Hunt

Overview

- Data Visualisation
- Drug Discovery Data
- Multi-Parameter Optimisation
- Using Data Visualisation to Drive Optimisation
- Conclusions

Using Data Visualisation

optibrium



Why use data visualisation?

Our visual system is *extremely* well built for visual analysis

- The optic nerve is a very big pipe
- Our brains are very good at edge detection, shape recognition and pattern matching

(Noah Iliinsky, Amazon Web Services, ComplexDiagrams.com)

BUT…

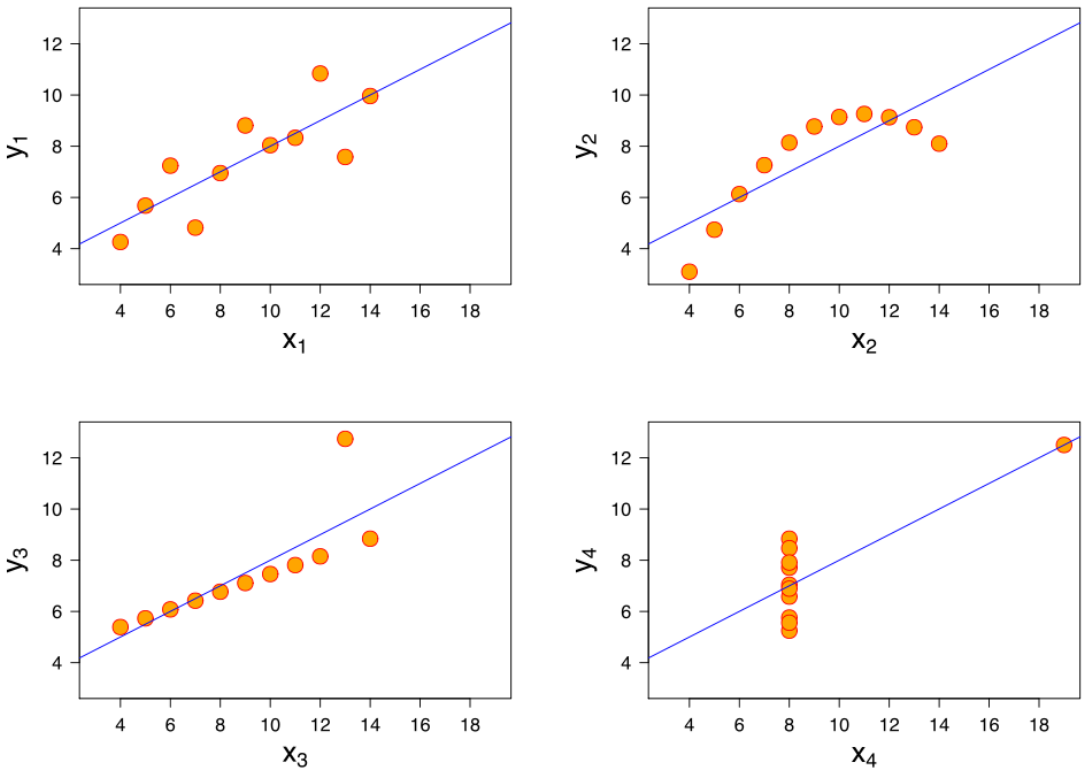
Data visualisation creates powerful, elegant images from complex data. It's like good prose: a pleasure to experience and a force for good in the right hands, but also seductive and potentially deceptive. … . Too much data visualisation is the statistical equivalent of dazzle camouflage: striking looks grab our attention but either fail to convey useful information or actively misdirect us. *(Tim Hartford – Financial Times)*

Use of data visualisation

When the statistics deceive us…

Anscombe's quartet

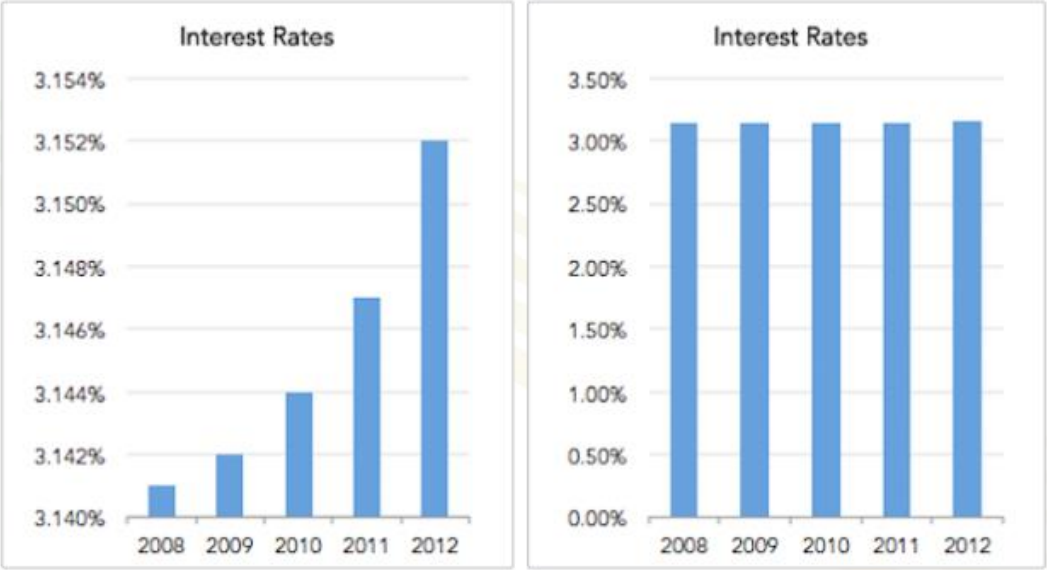
Property	Value
Mean(x)	9
Variance(x)	11
Mean(y)	7.5 (2dp)
Variance(y)	4.122 or 4.127 (3dp)
Pearson r	0.816 (3dp)
Linear regression	y = 3 + 0.5x (2dp)

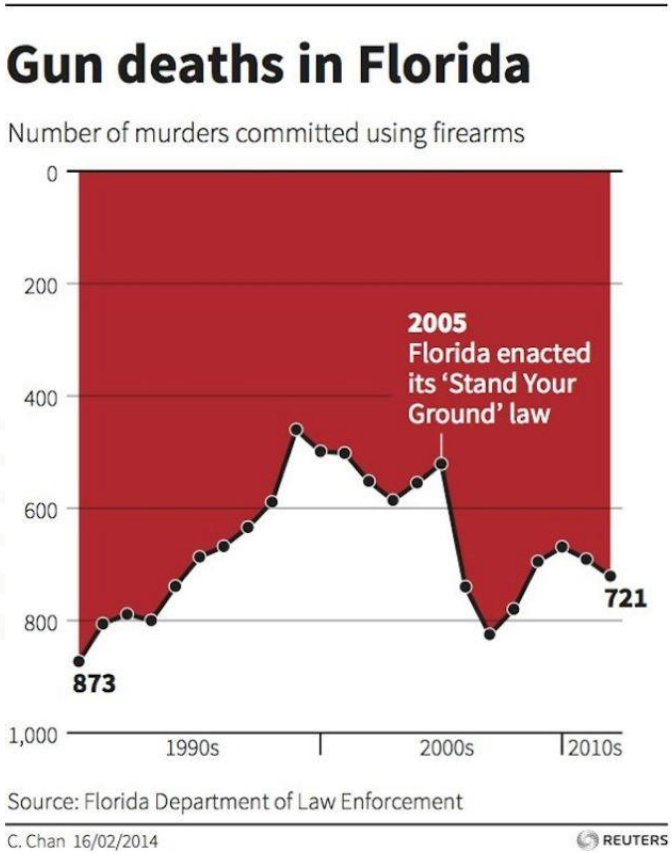


en.wikipedia.org/wiki/Anscombe%27s_quartet

Misuse of data visualisation

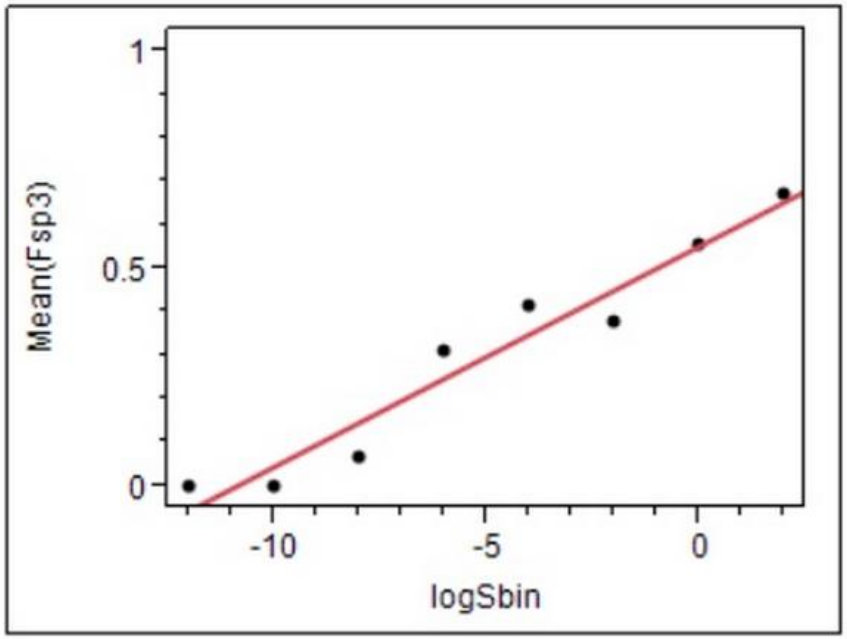
Same Data, Different Y-Axis



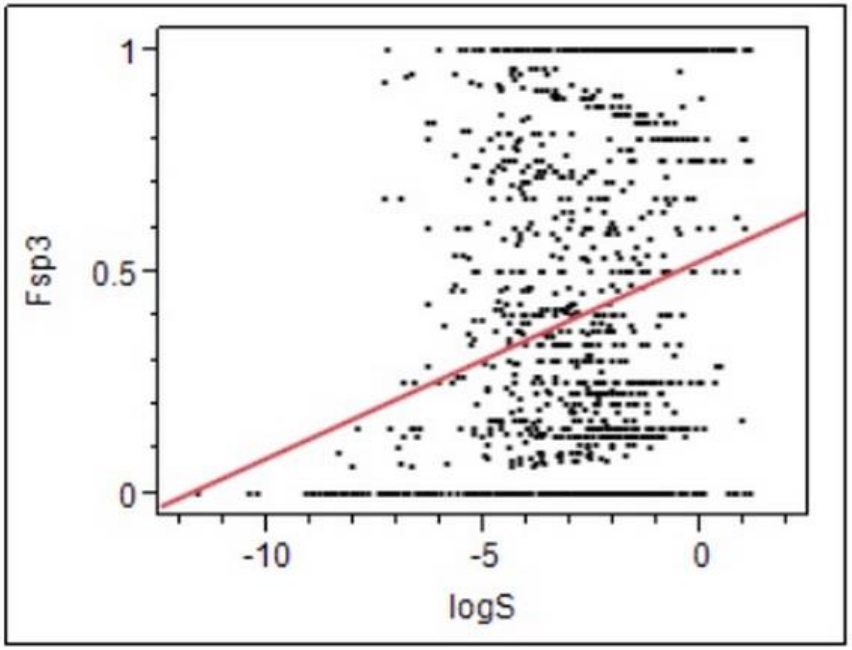


Ravi Parikh - Heap Analytics:
gizmodo.com/how-to-lie-with-data-visualization-1563576606

Misuse of data visualisation

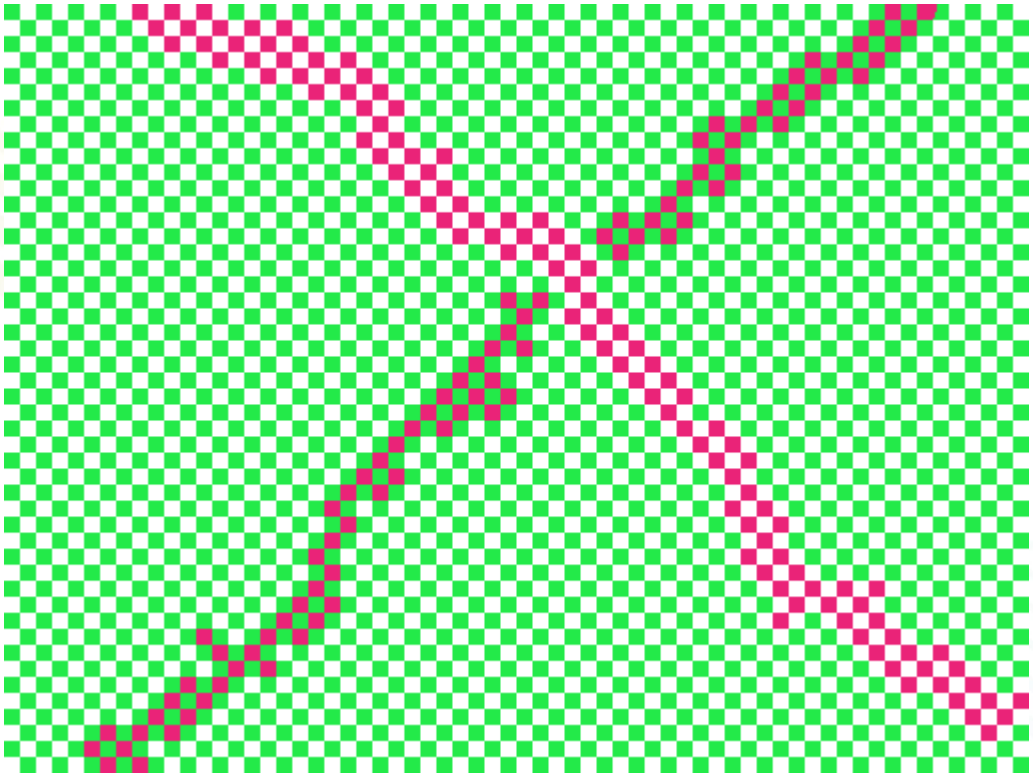


From binning this data…



Inflation of correlation in the pursuit of drug-likeness.
Kenny, Montanari - J Comput Aided Mol Des. 2013 Jan;27(1):1-13

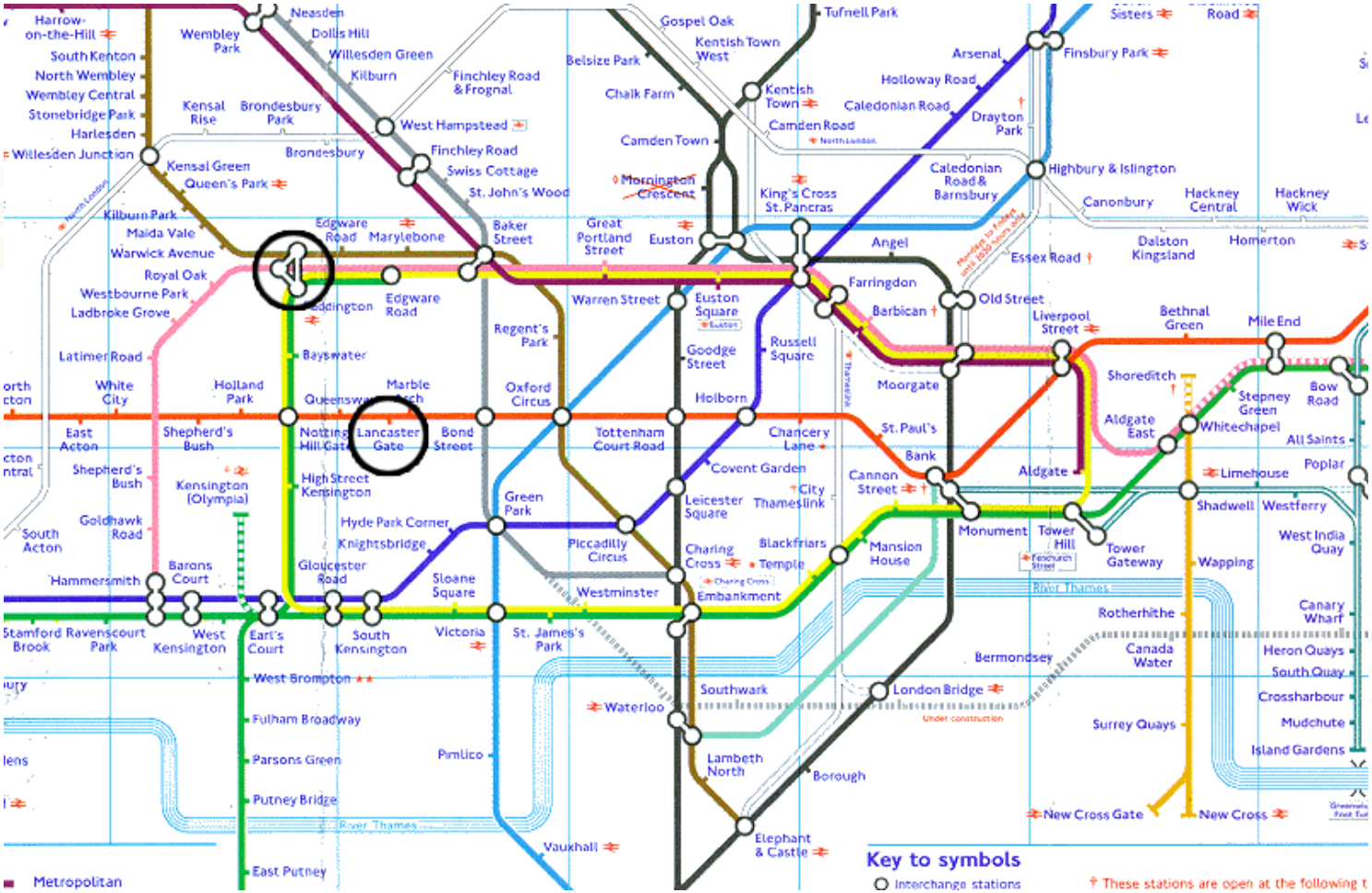
Colour can have the wrong impact…



www.grand-illusions.com/opticalillusions/square/

Misuse of data visualisation

Using a good visualisation for the wrong purposes…

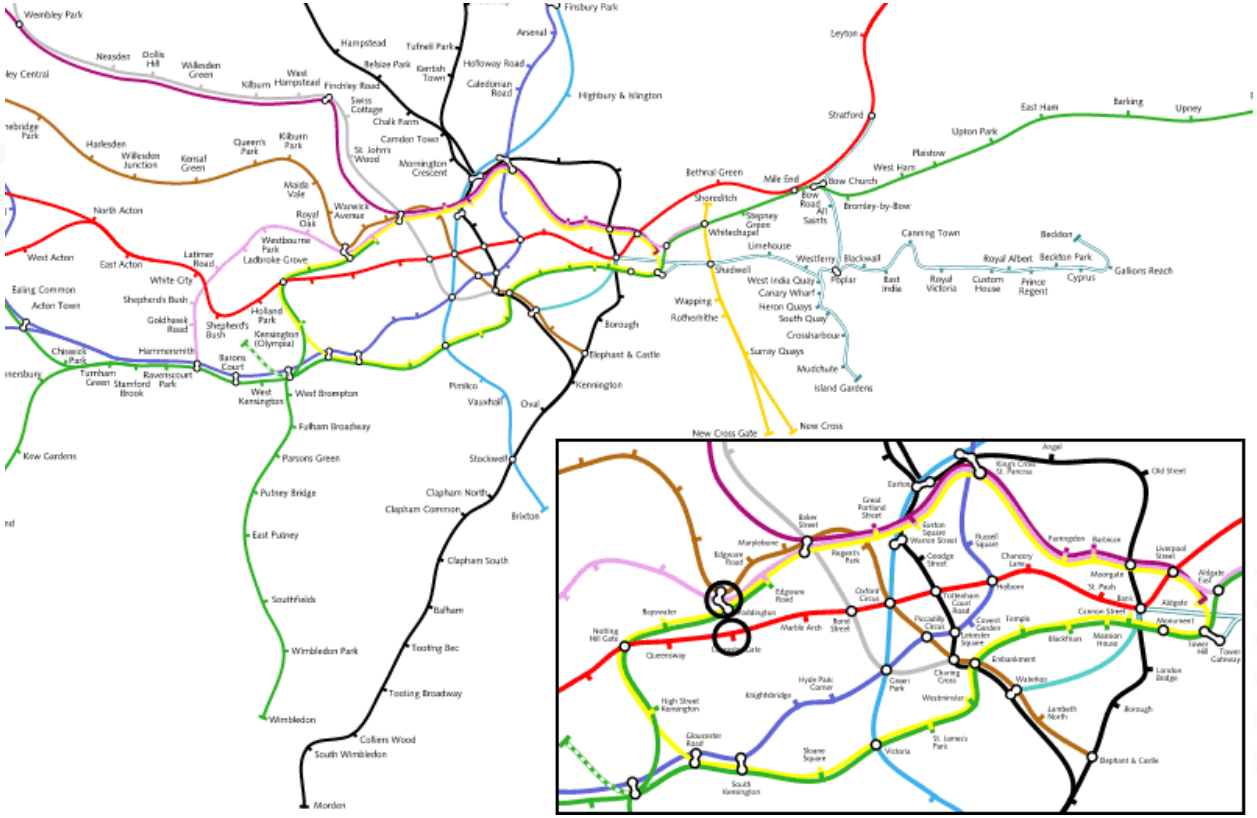


Sayf Sharif

lunametrics.com/blog/2013/02/04/power-danger-data-visualization

Misuse of data visualisation

Using a good visualisation for the wrong purposes…



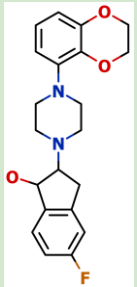
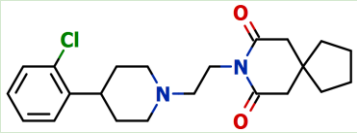
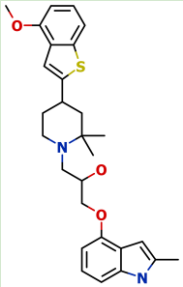
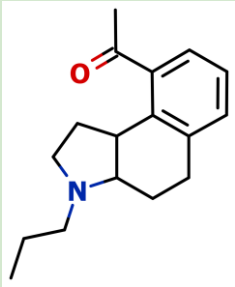
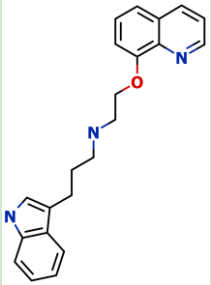
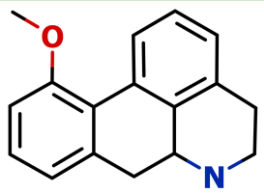
Sayf Sharif

lunametrics.com/blog/2013/02/04/power-danger-data-visualization

Back to drug discovery…

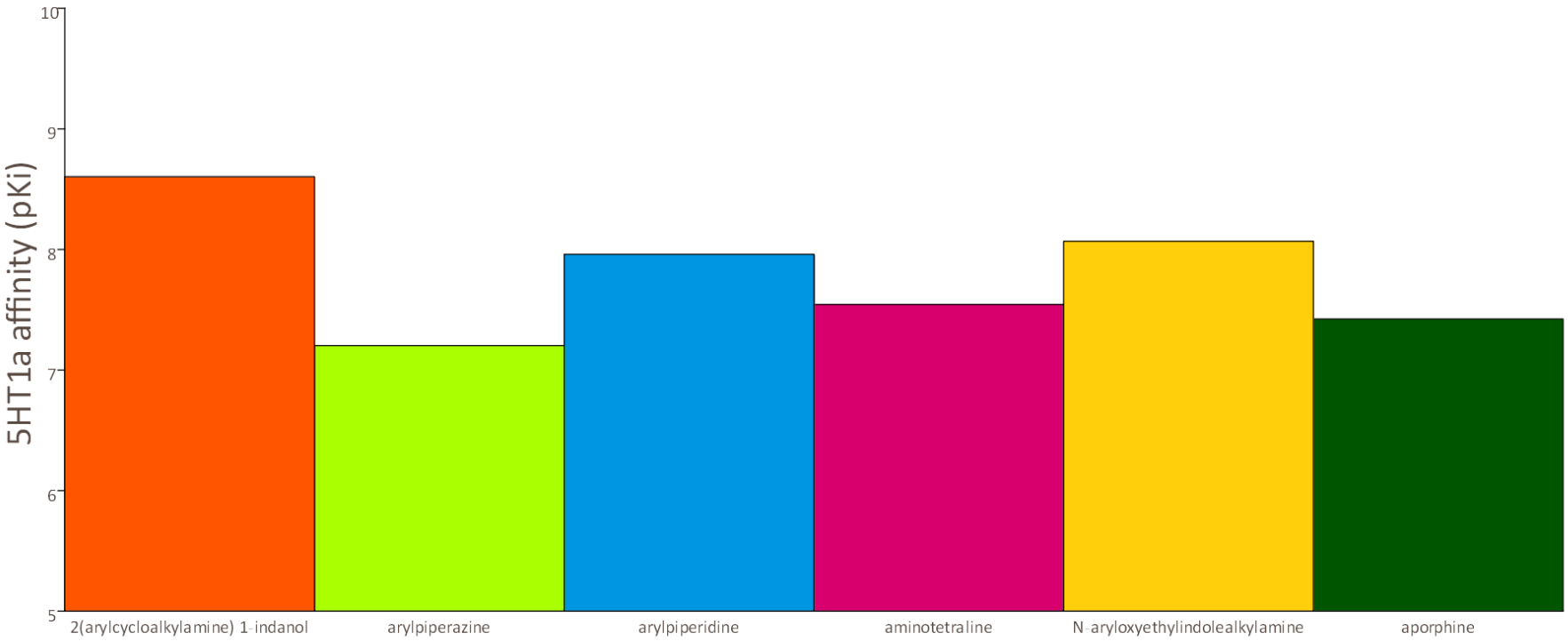
Let's looks at some drug discovery data:

- Library of 264 5HT1a compounds
- Measured potencies and other ADME/physicochemical properties
- Six different chemotypes:

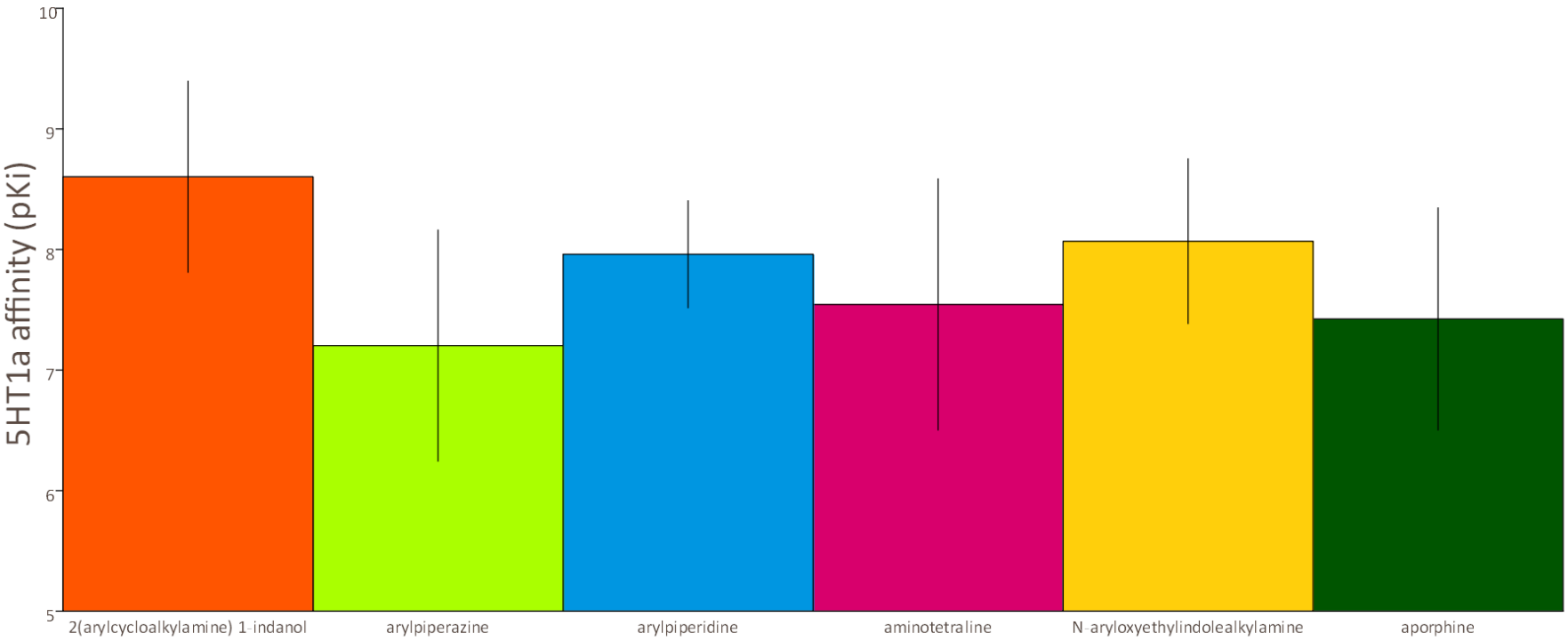
2(arylcycloalkylamine) 1-indanols (27)	Arylpiperazines (120)	Arylpiperidines (17)	Aminotetralines (51)	N-aryloxyethylindolealkylamines (29)	Aporphines (20)
					

Let's think about how we might prioritise these…

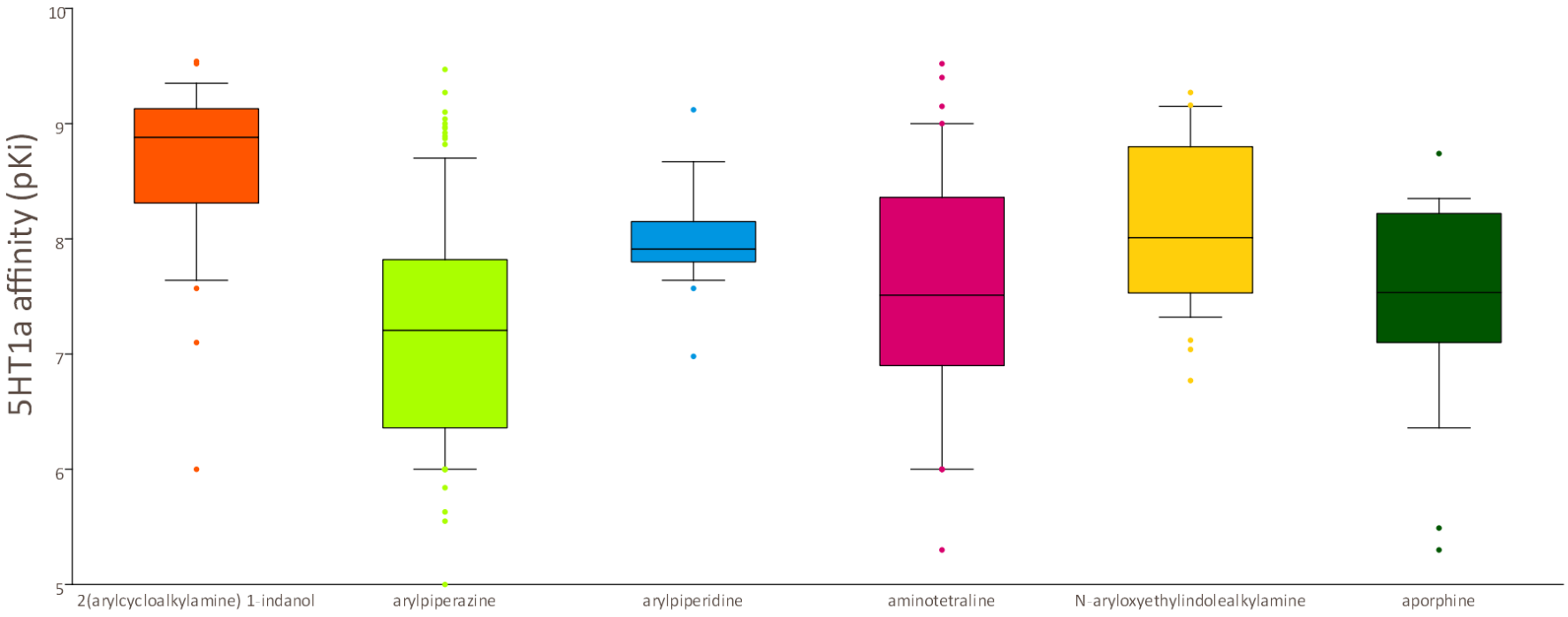
Let's start with potency



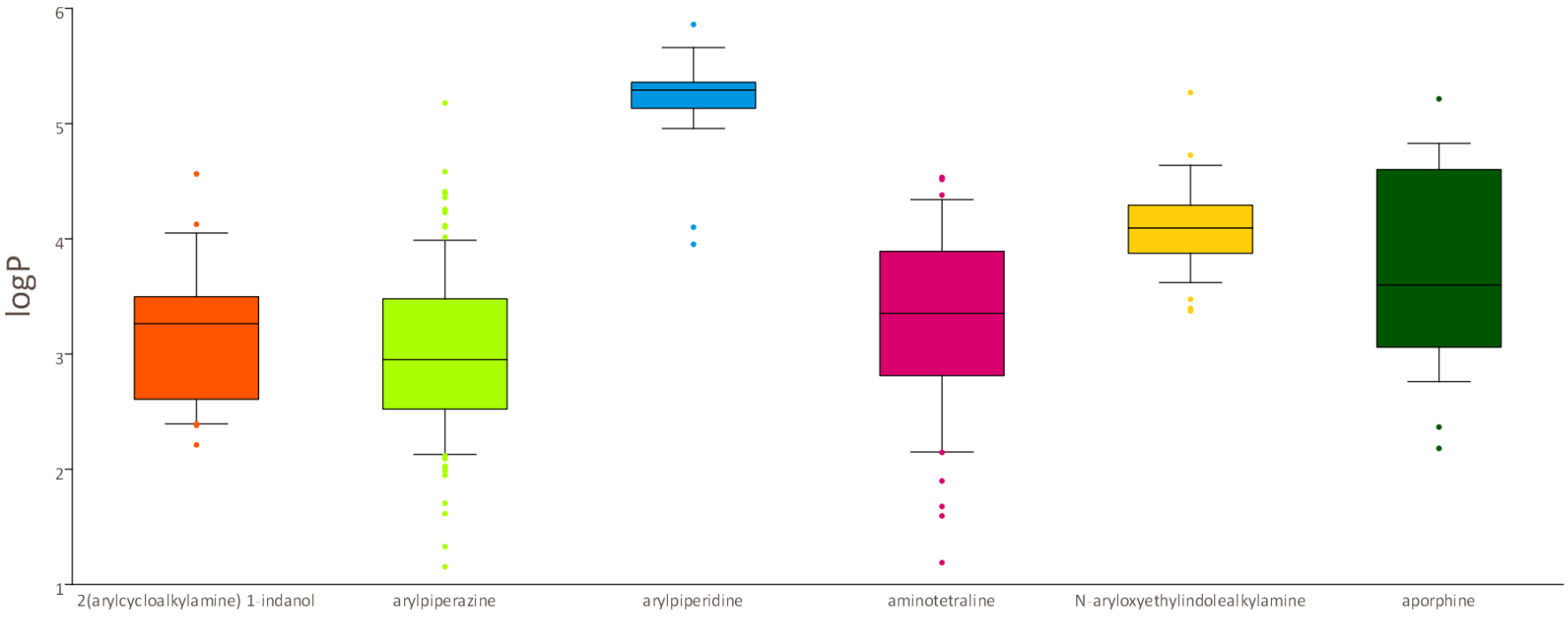
Let's start with potency – with error bars



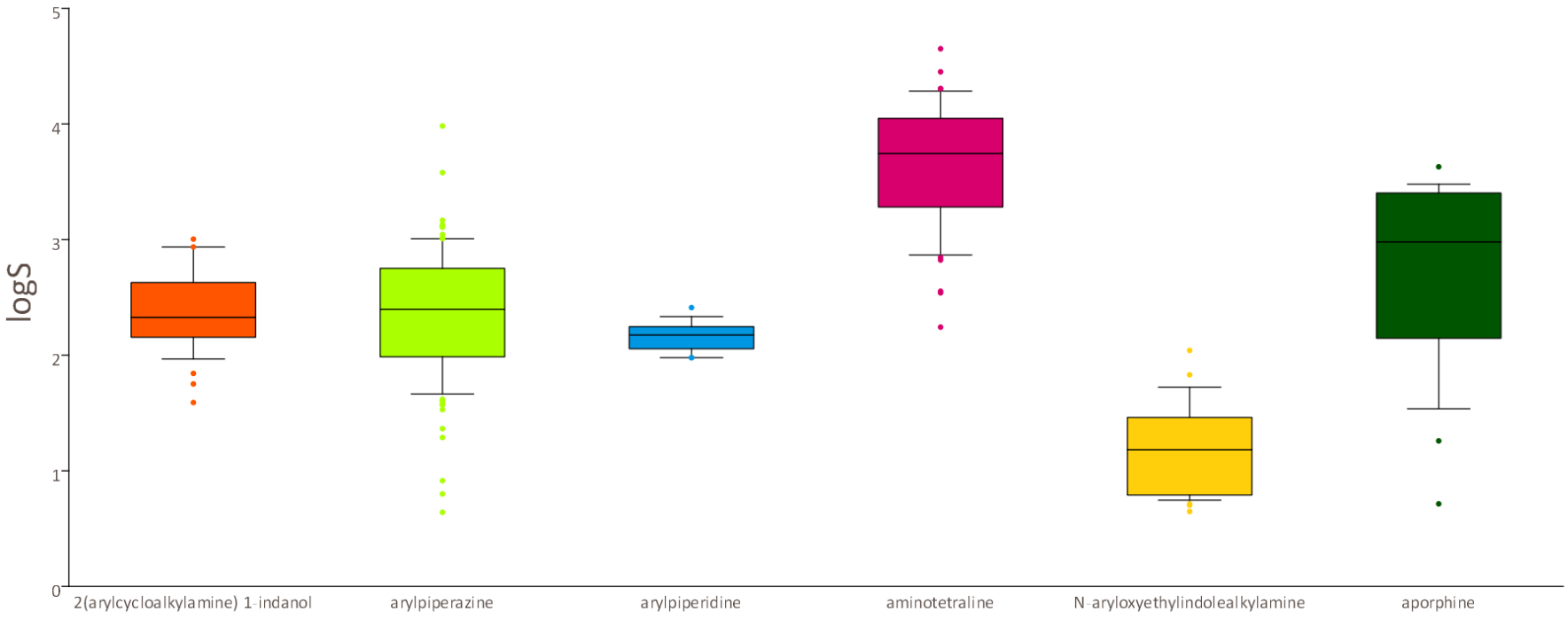
Box plot gives a clearer picture



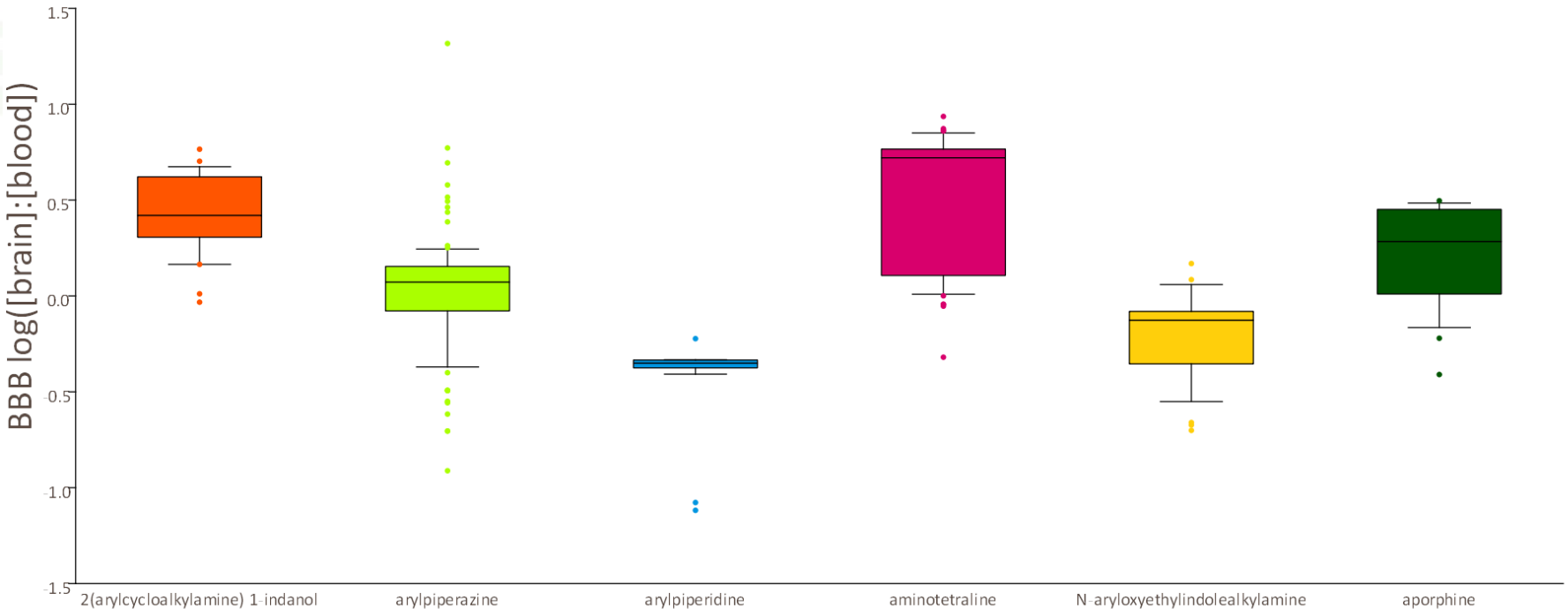
Other properties: logP



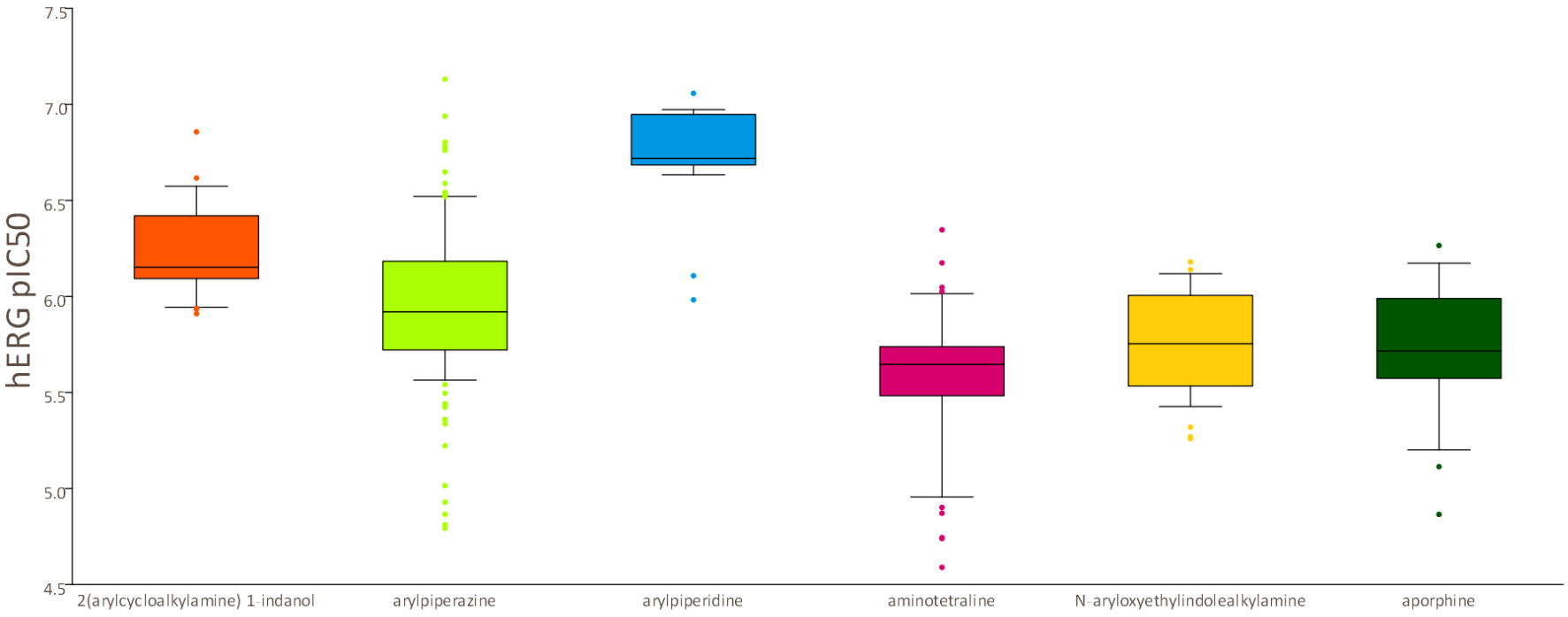
Other properties: Solubility (logS)



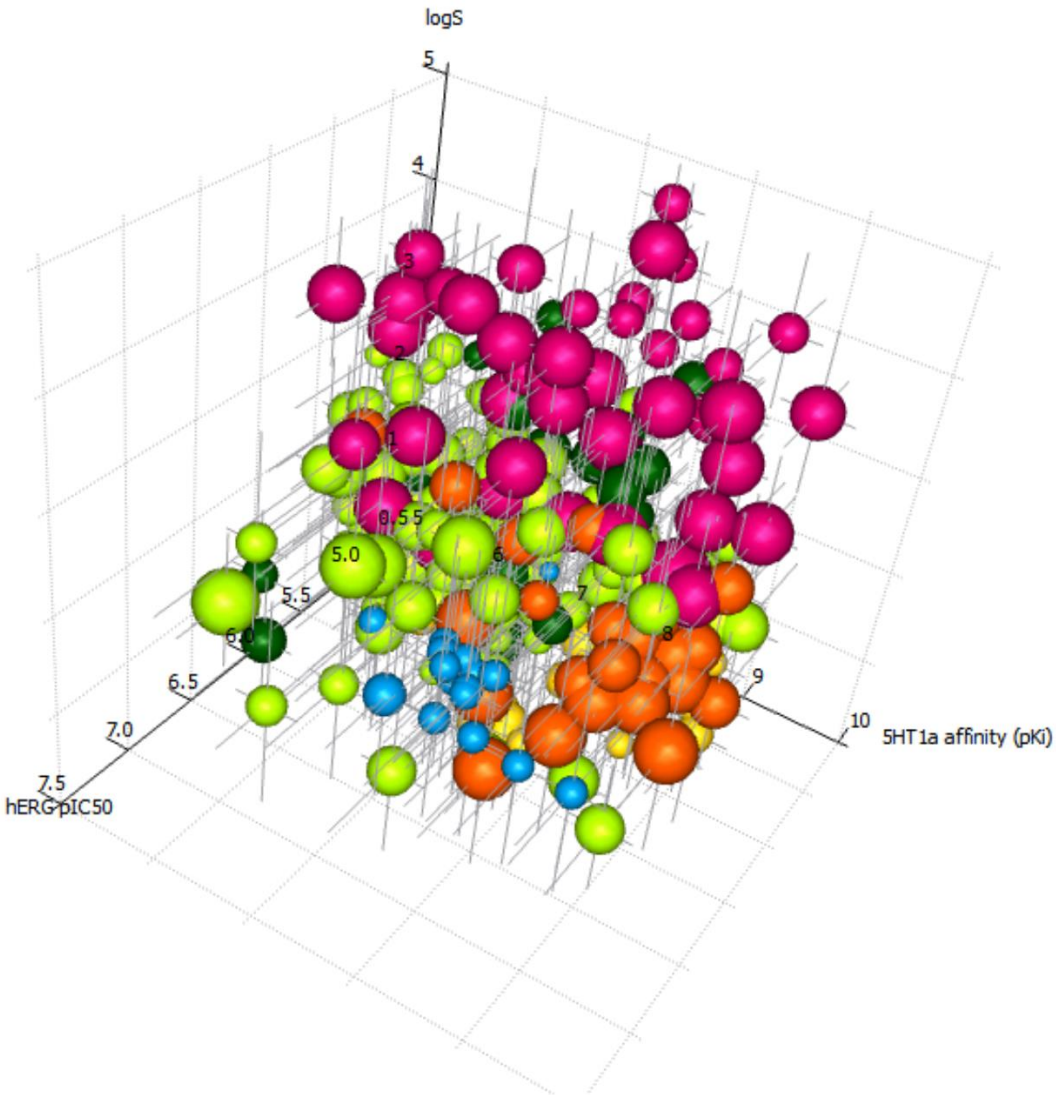
Other properties: BBB penetration



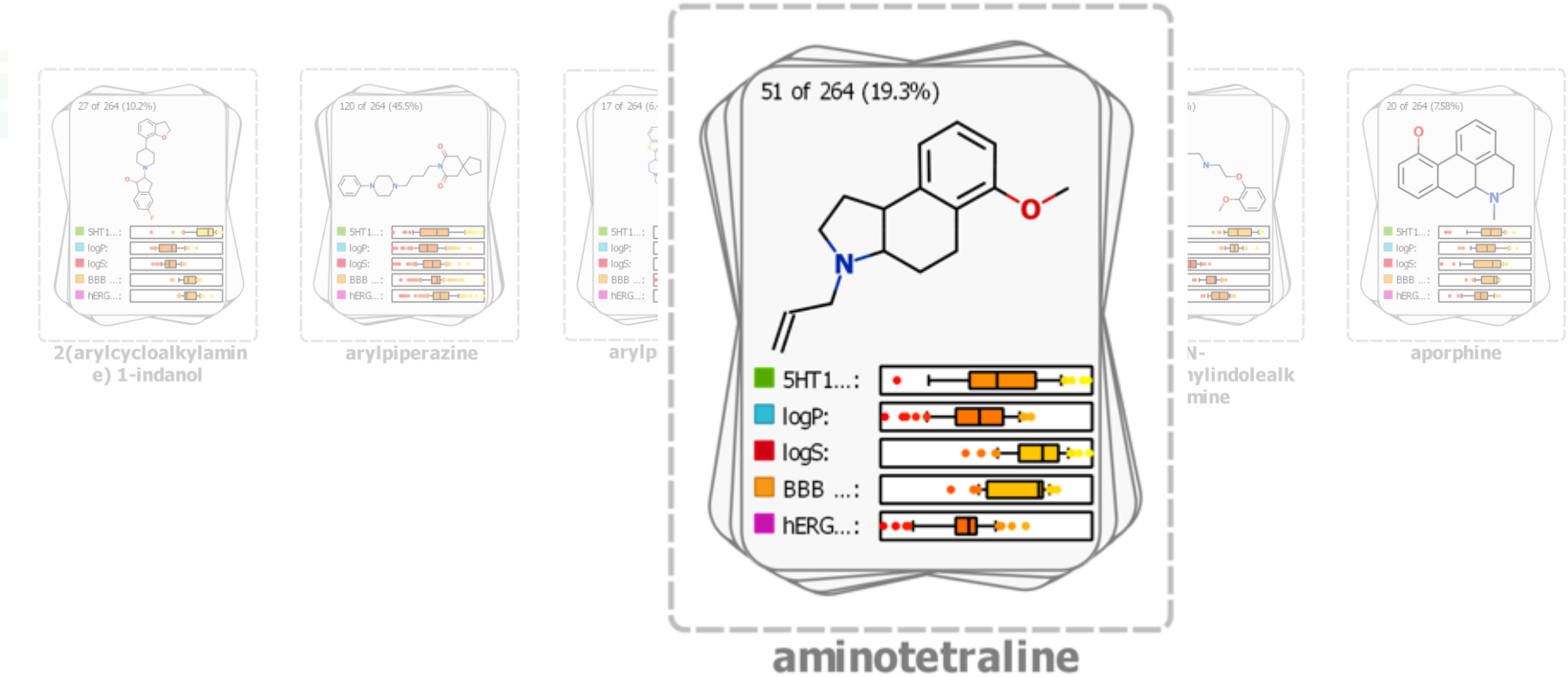
Other properties: hERG pIC50



Seeing them all together?



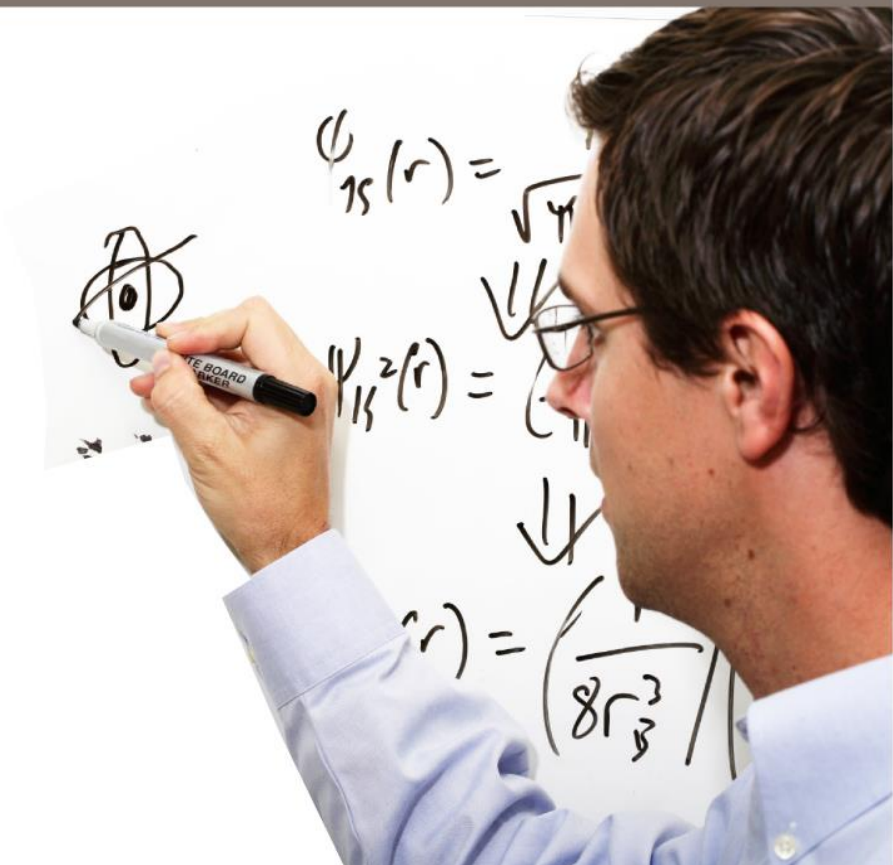
Seeing them all together



Card View™

Drug Discovery Data

optibrium



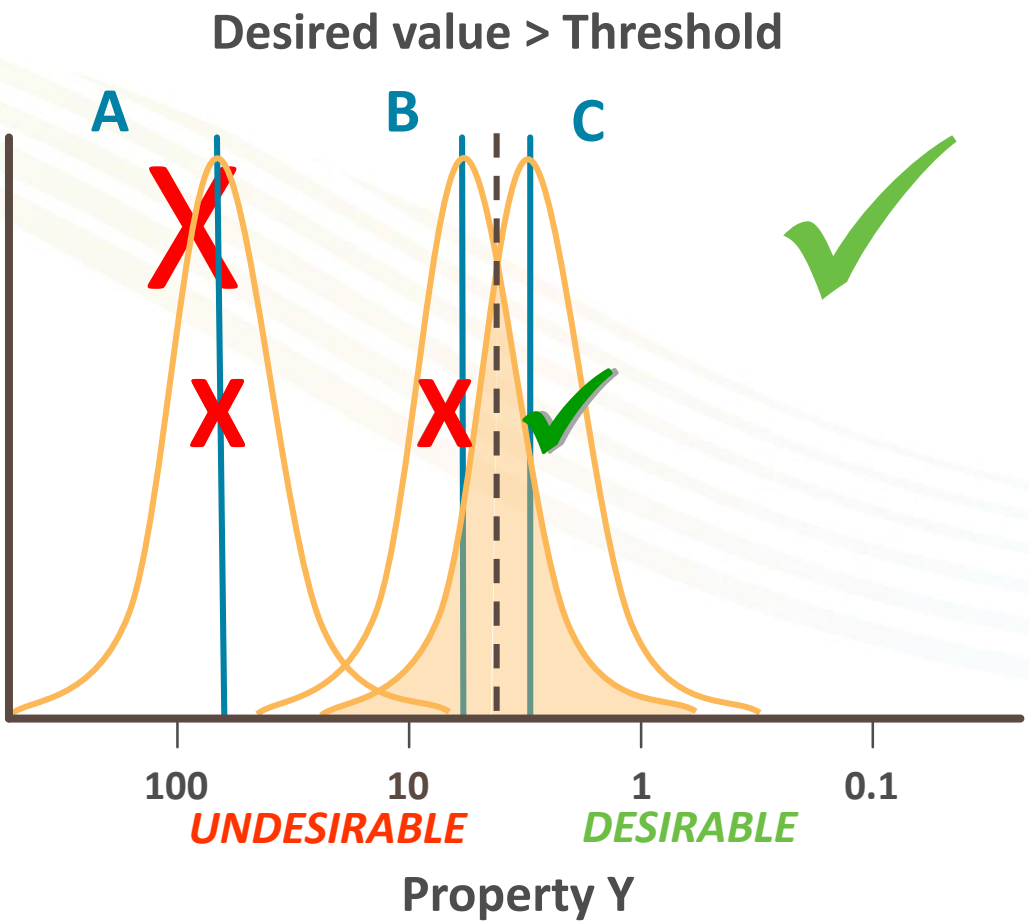
Data in drug discovery

- What's certain?
  - We know some simple properties of our compounds
- What's not so certain?
  - *In vitro*/*In vivo* measurements
    - experimental variability
  - *In silico* predictions
    - statistical error
  - Inference/translation

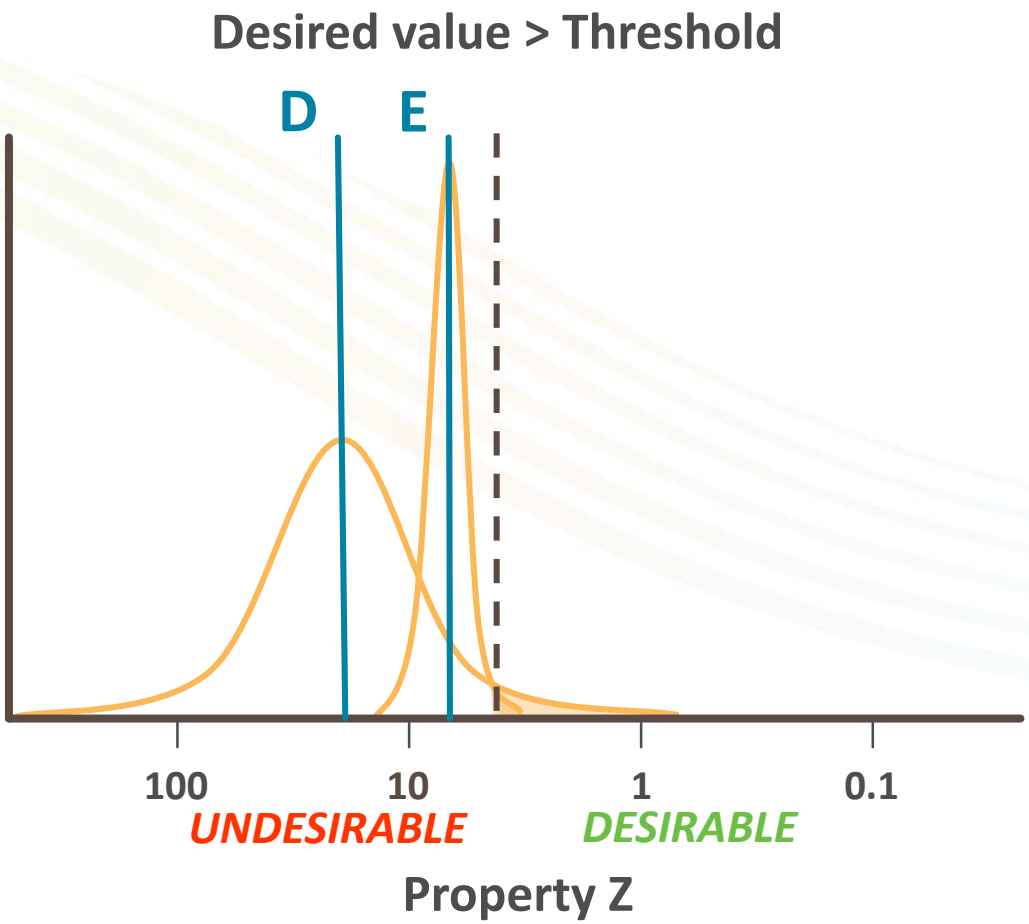
The Challenges:
Uncertain data

- So what does that mean…
- A good RMSE for logS (solubility) is 0.6
- Assuming normal distribution this means that when we have a logS value of 2 (that's 100µM) then
  - 68% of the time this represents an actual value between 1.4 and 2.6 (25µM to 400µM)
  - 95% of the time this represents an actual value between 0.8 and 3.2 (6µM to 1.6mM)
  - 99% of the time this represents an actual value between 0.2 and 3.8 (1.6µM to 6.3mM)

Importance of Uncertainty



Importance of Uncertainty

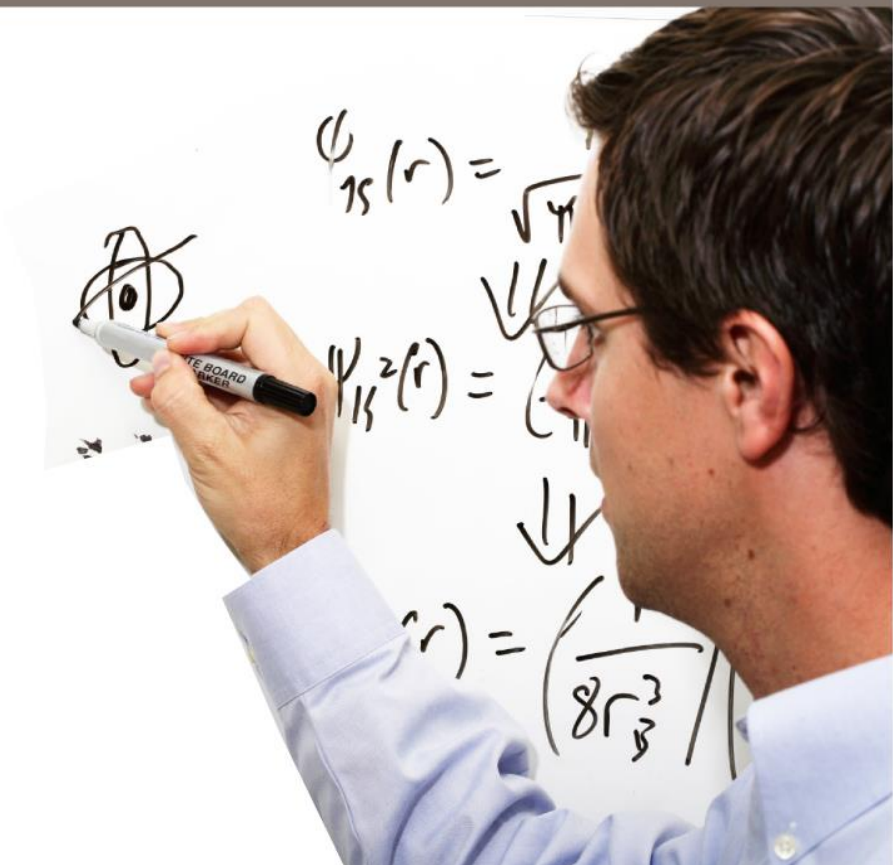


The Challenges:
…and one more thing

- We probably have quite a few properties we need to optimise!
  - Each will have their own uncertainty
  - Each will have its own criteria we'd like to achieve
  - Each will have its own level of importance relative to the other properties

Multi-Parameter Optimisation

optibrium



Back to our 5HT1a library

- Example criteria we might like to achieve for an ideal compound

Property	Desired value	Importance
Potency (pK_i)	> 7	High
logS (log µM)	> 1	High
Human Intestinal Absorption (category)	+	High
BBB log([brain]:[blood])	-0.2 -> 1	High
logP	0 -> 3.5	Medium
P-gp (category)	No	Medium
hERG pIC50	≤ 5	Medium
2C9 pK_i	≤ 6	Low
2D6 affinity (category)	Low/Medium	Low
Plasma protein binding (category)	Low	Low

Putting it all together (MPO):
Probabilistic Scoring* – Scoring Profile

Property	Desired Value	Importance
5HT1a affinity (pKi)	> 7	
logS	> 1	
HIA category	+	
BBB log([brain]:[blood])	-0.2 -> 1	
logP	0 -> 3.5	
P-gp category	no	
hERG pIC50	⩽ 5	
2C9 pKi	⩽ 6	
2D6 affinity category	low medium	
PPB90 category	low	

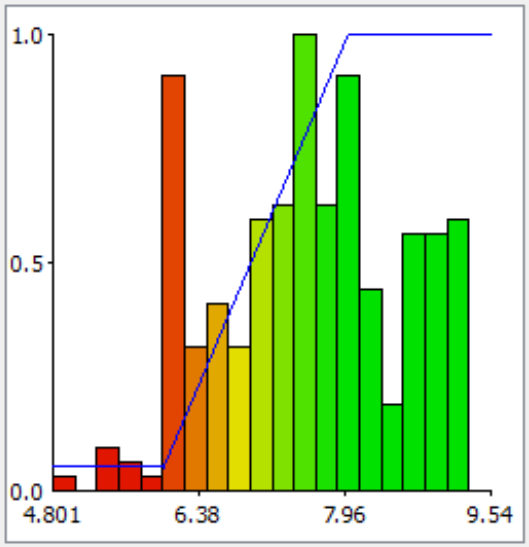
Score Editor: 5HT1a affinity (pKi)

Category:

Importance:

Insert | Delete

Range	Score
-inf -> 6	0.05
6 -> 8	0.05 -> 1.00
8 -> inf	1.00



OK | Cancel

* Segall *et al.* (2009) Chem. & Biodiv. **6** p. 2144

Multi-parameter Optimisation
Probabilistic Scoring*

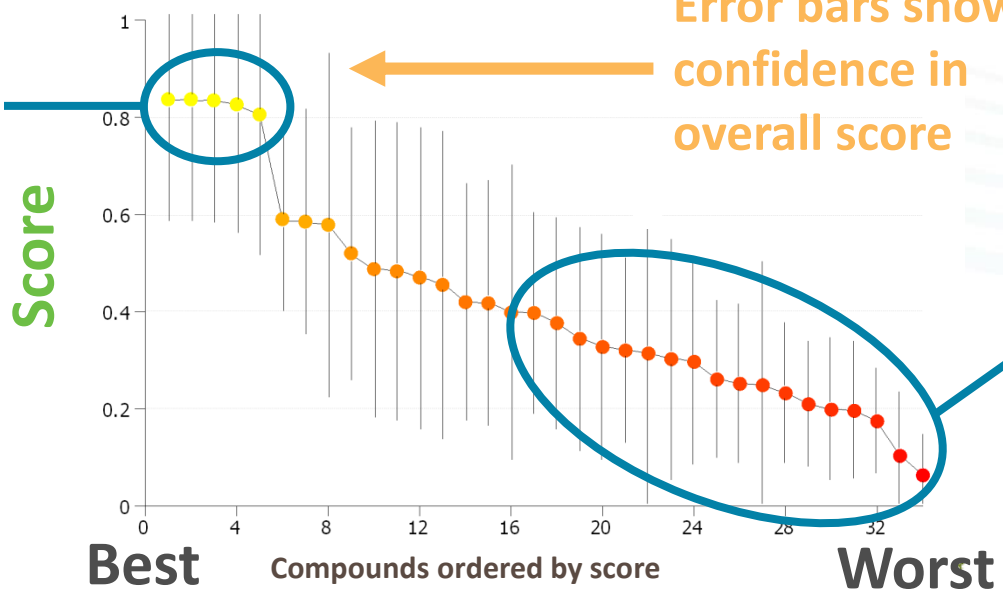
- **Property data**
  - Experimental or predicted
- **Criteria for success**
  - Relative importance
- **Uncertainties in data**
  - Experimental or statistical

- **Score (Likelihood of Success)**
- **Confidence in score**

Data do not separate these as error bars overlap

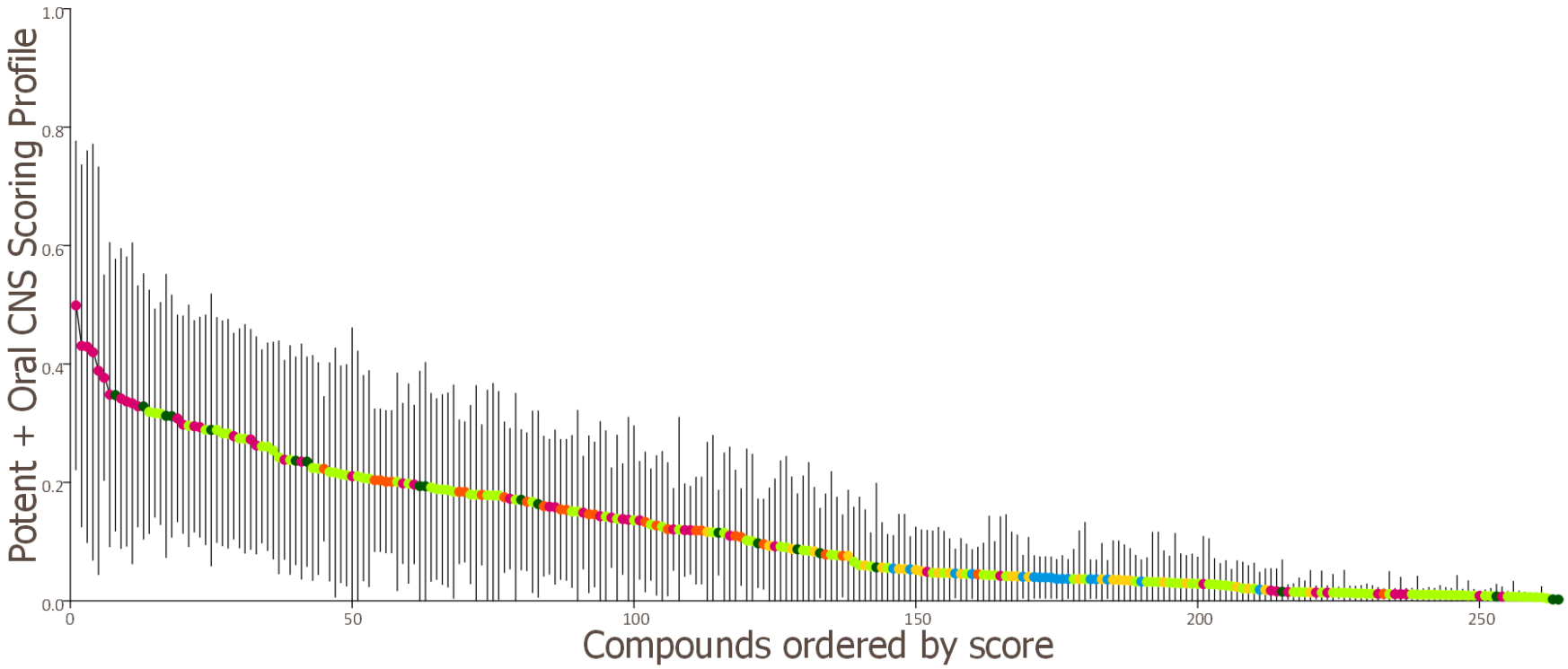
Error bars show confidence in overall score

Bottom 50% may be rejected with confidence

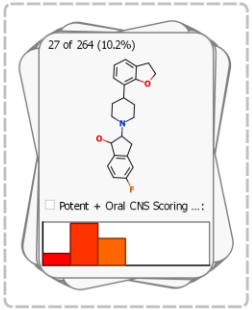


*M.D. Segall (2012) Curr. Pharm. Des. **18**(9) pp. 1292-1310

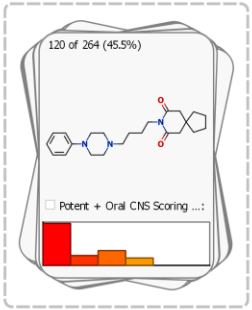
Snake plot for complete library



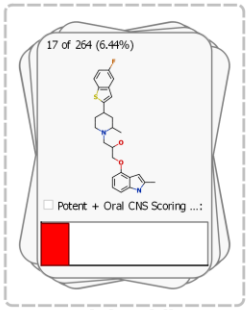
Stacks for each chemistry type show distribution of scores



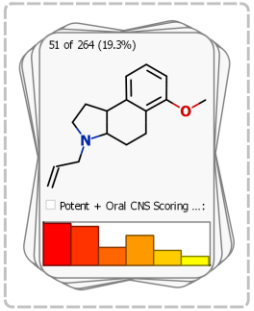
2(arylcycloalkylamine) 1-indanol



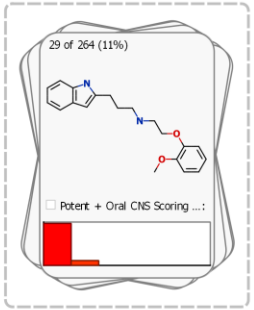
arylpiperazine



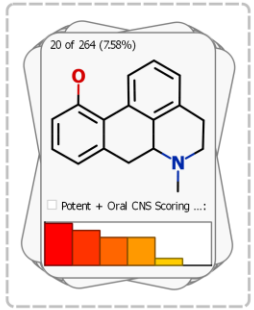
arylpiperidine



aminotetraline

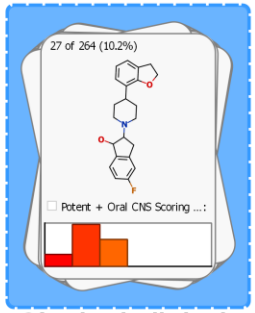


N-aryloxyethylindolealkylamine

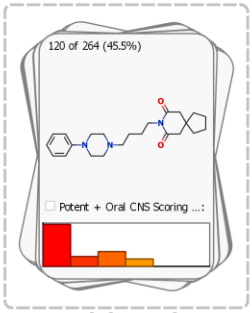


aporphine

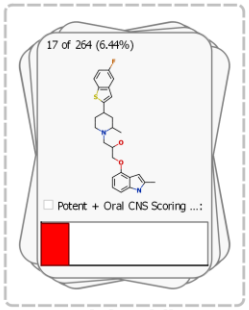
Each chemistry in turn



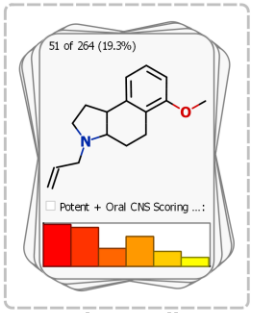
2(arylcycloalkylamine) 1-indanol



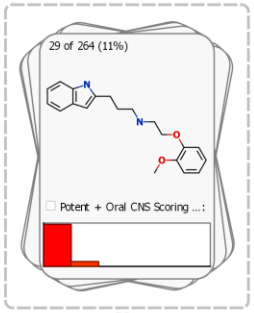
arylpiperazine



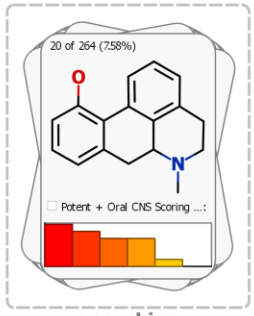
arylpiperidine



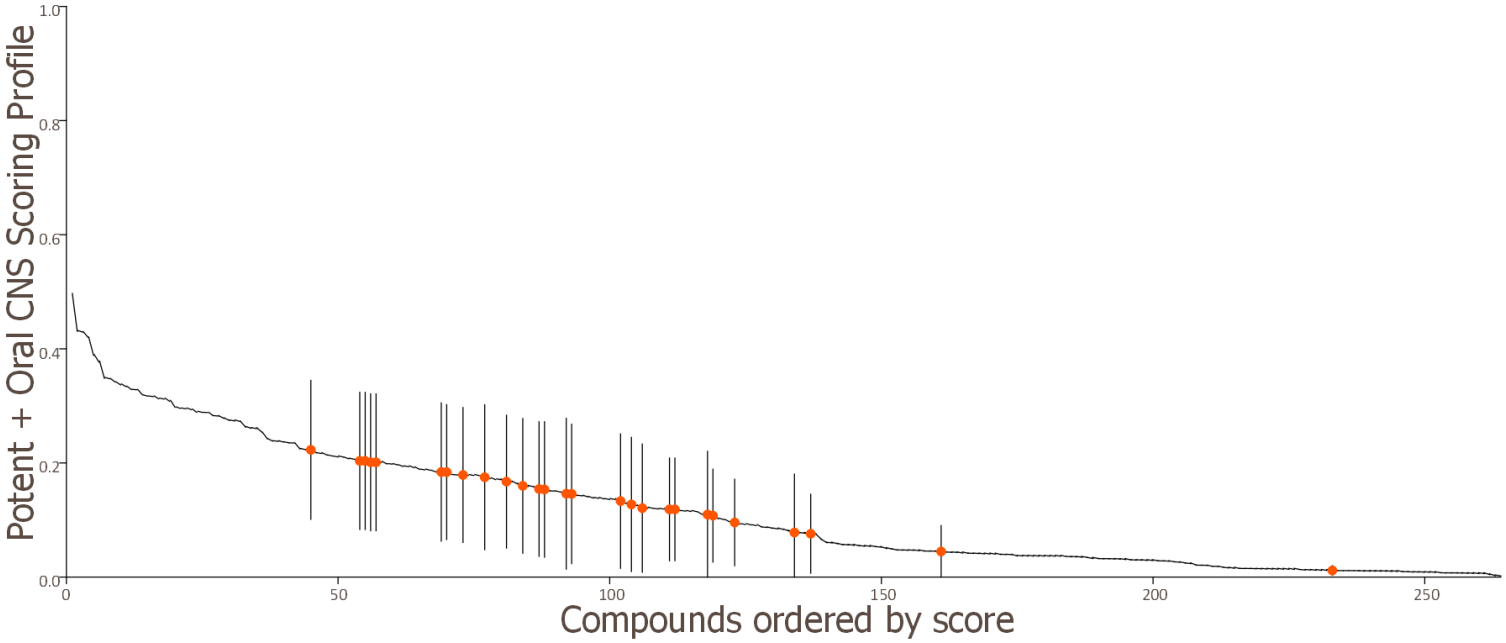
aminotetraline



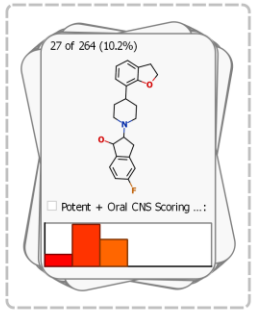
N-aryloxyethylindolealkylamine



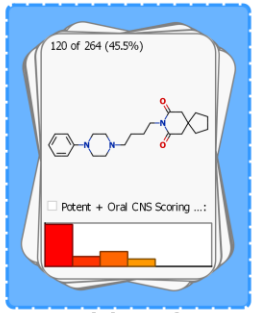
aporphine



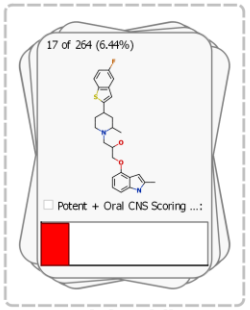
Each chemistry in turn



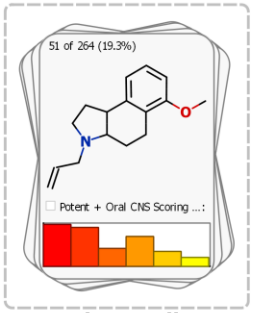
2(arylcycloalkylamine) 1-indanol



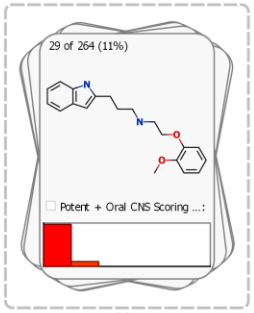
arylpiperazine



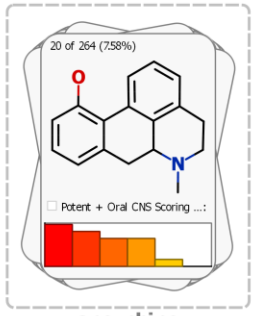
arylpiperidine



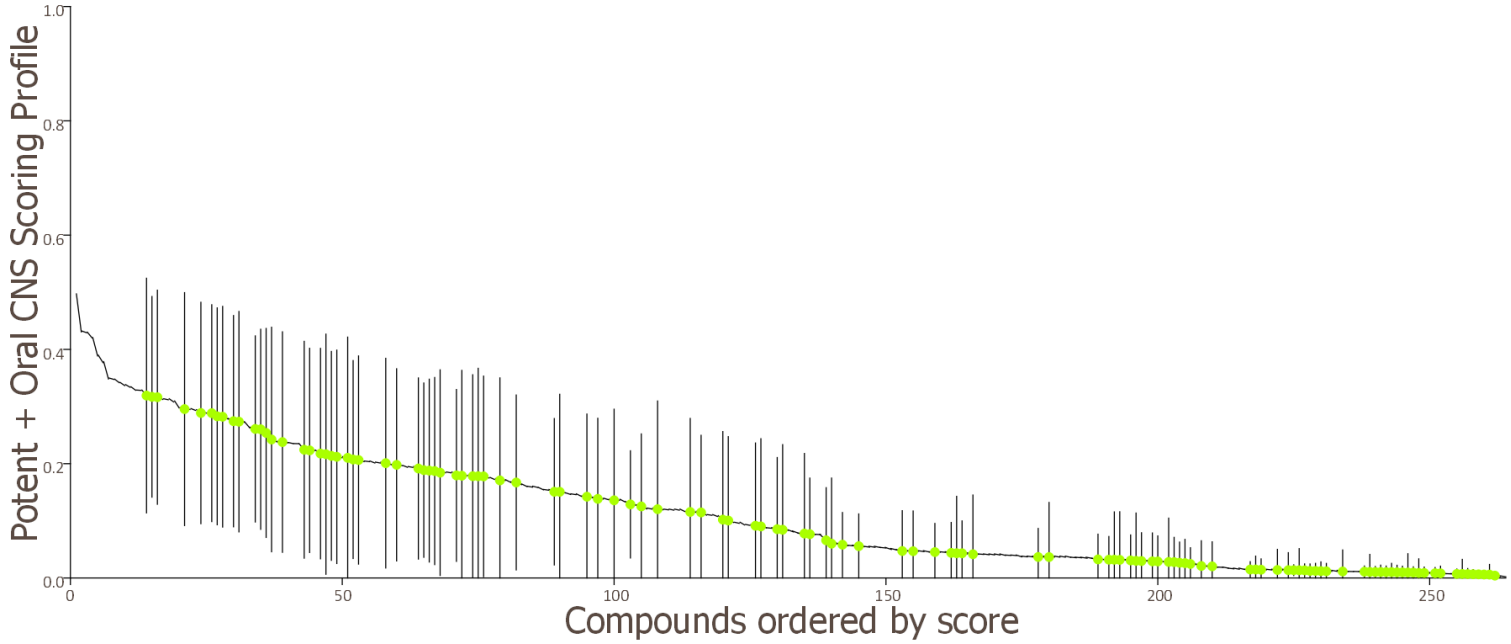
aminotetraline



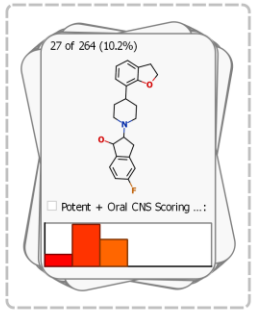
N-aryloxyethylindolealkylamine



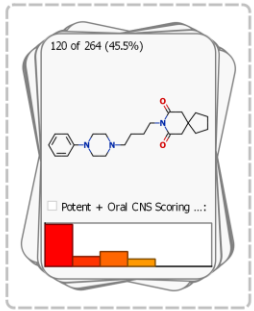
aporphine



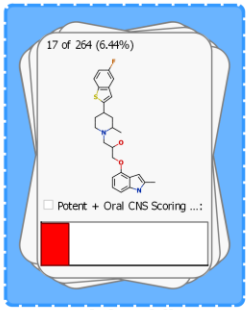
Each chemistry in turn



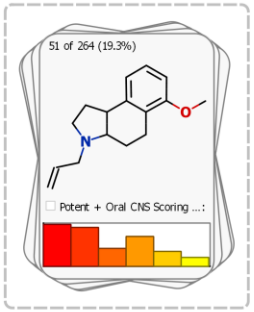
2(arylcycloalkylamine) 1-indanol



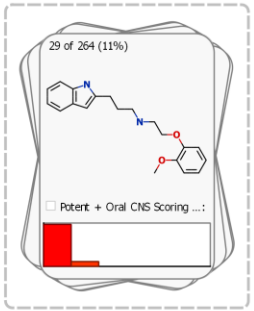
arylpiperazine



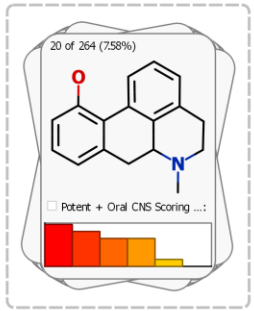
arylpiperidine



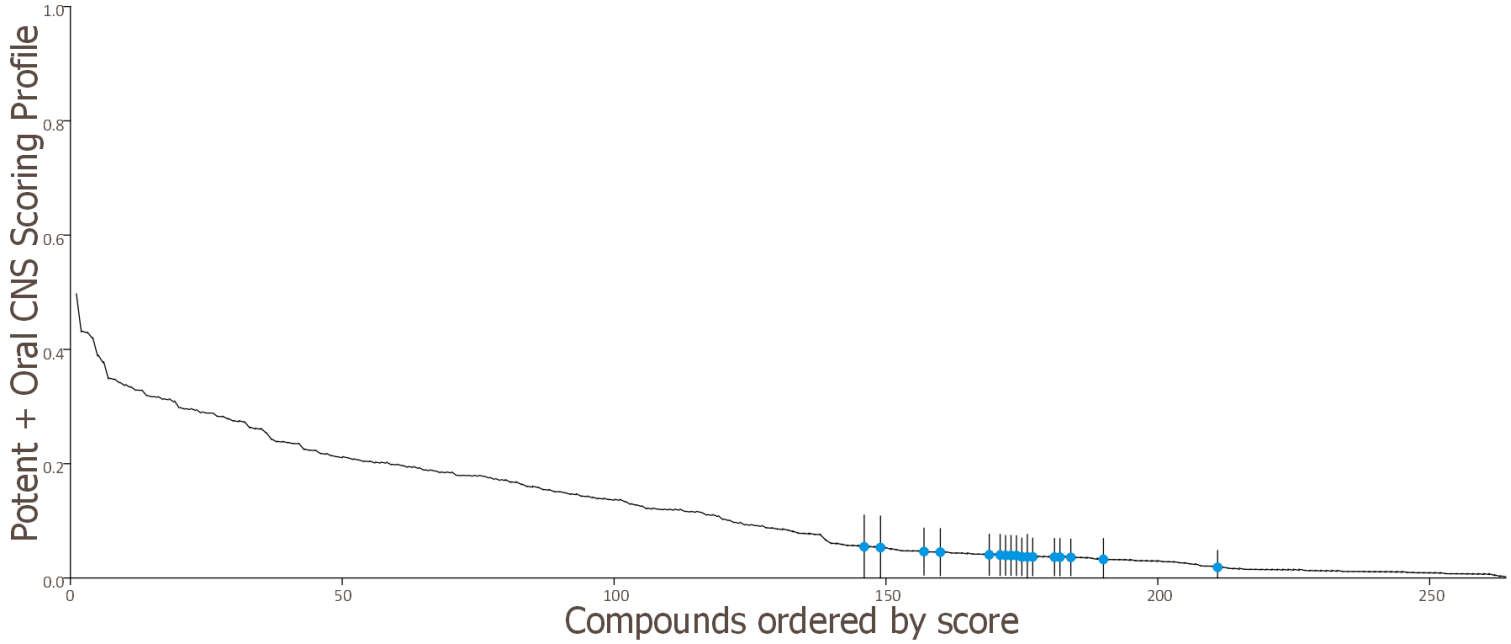
aminotetraline



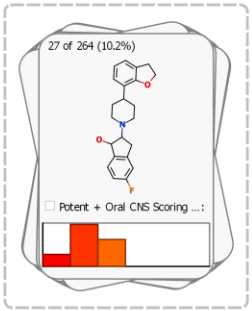
N-aryloxyethylindolealkylamine



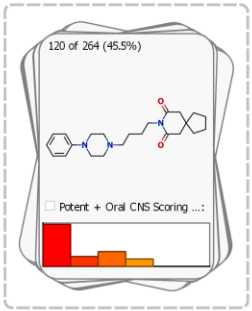
aporphine



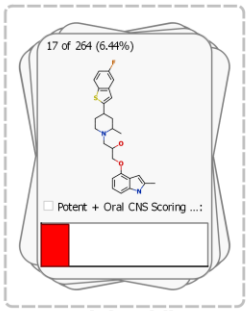
Each chemistry in turn



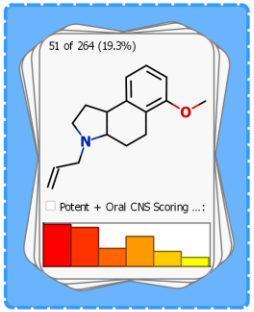
2(arylcycloalkylamine) 1-indanol



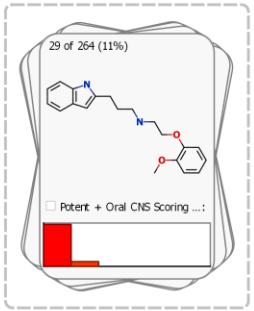
arylpiperazine



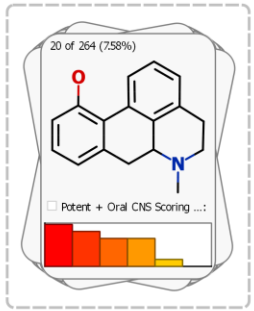
arylpiperidine



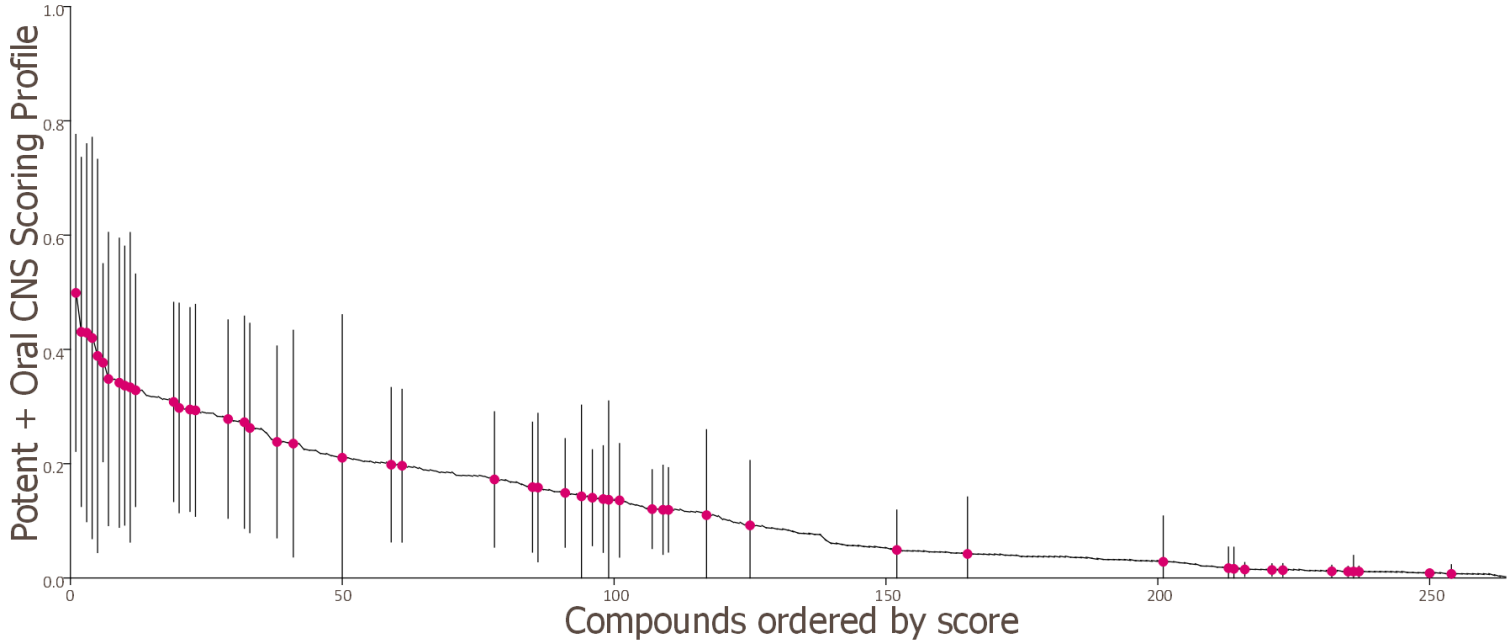
aminotetraline



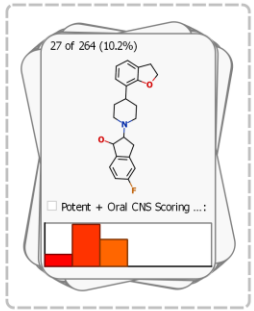
N-aryloxyethylindolealkylamine



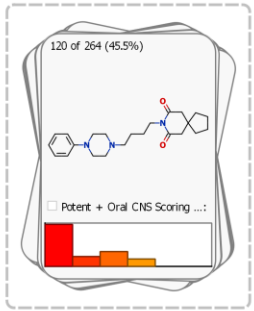
aporphine



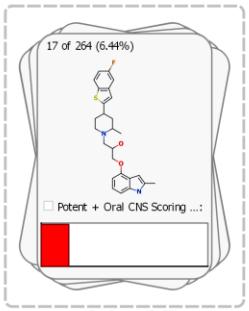
Each chemistry in turn



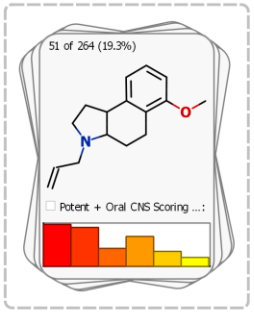
2(arylcycloalkylamine) 1-indanol



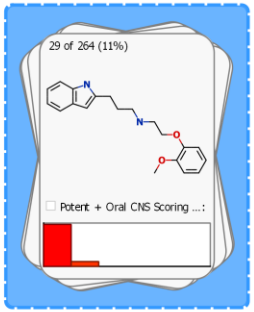
arylpiperazine



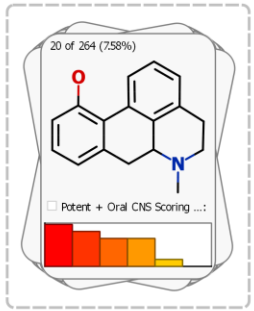
arylpiperidine



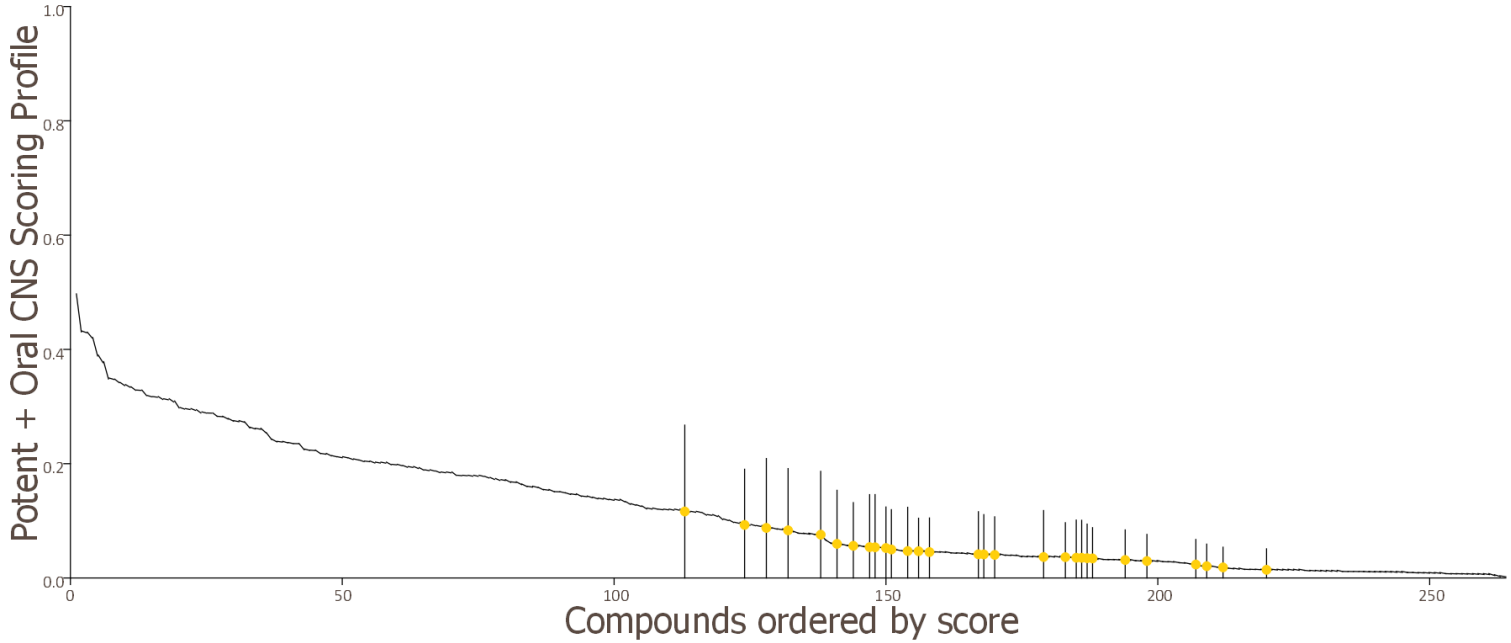
aminotetraline



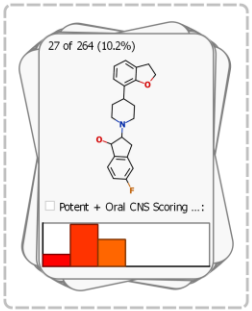
N-aryloxyethylindolealkylamine



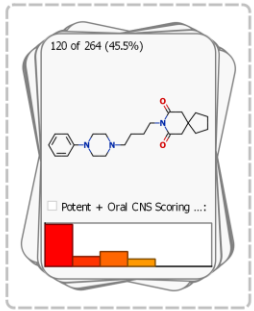
aporphine



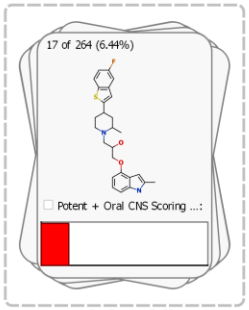
Each chemistry in turn



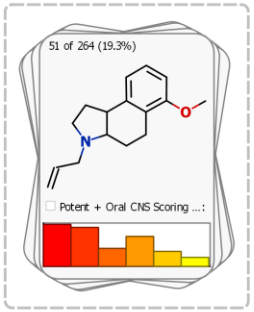
2(arylcycloalkylamine) 1-indanol



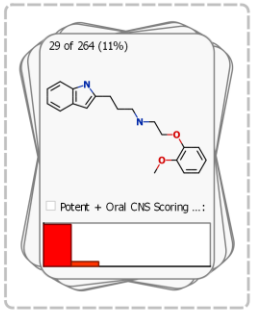
arylpiperazine



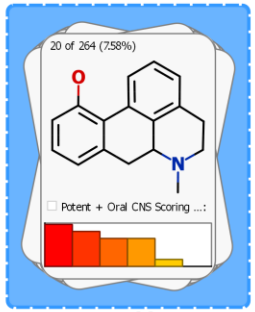
arylpiperidine



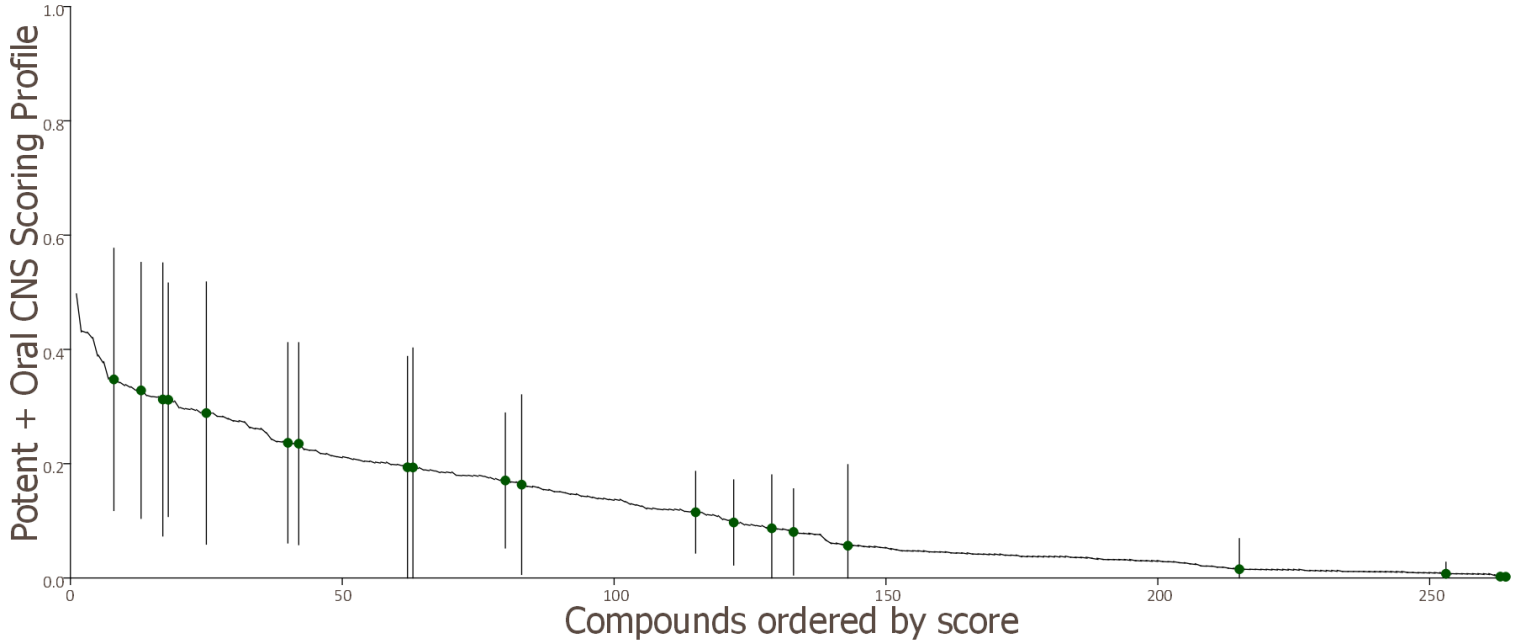
aminotetraline



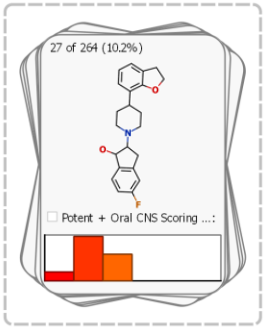
N-aryloxyethylindolealkylamine



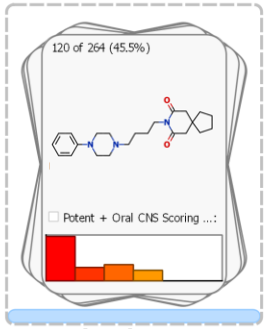
aporphine



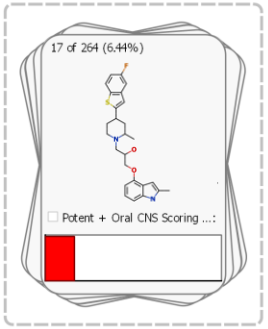
An appropriate selection?



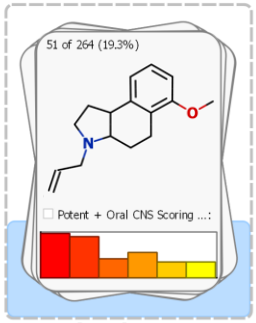
Chemistry = 2(arylcycloalkylamine) 1-indanol



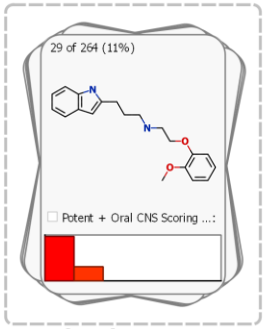
Chemistry = arylpiperazine



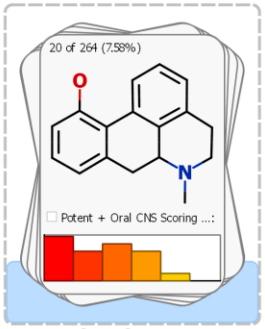
Chemistry = arylpiperidine



Chemistry = aminotetraline



Chemistry = N-aryloxyethylindolealkylamines

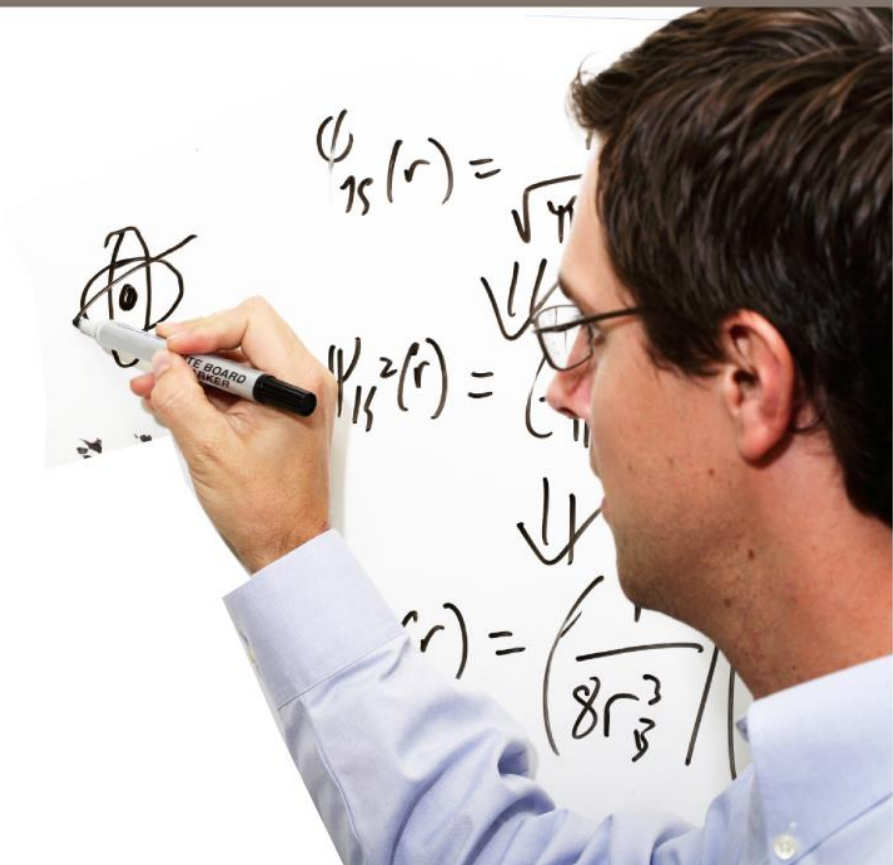


Chemistry = aporphine

(N.B. There are 2.77×10^{54} possible ways to select 50 compounds)

Using Data Visualisation to Drive Optimisation

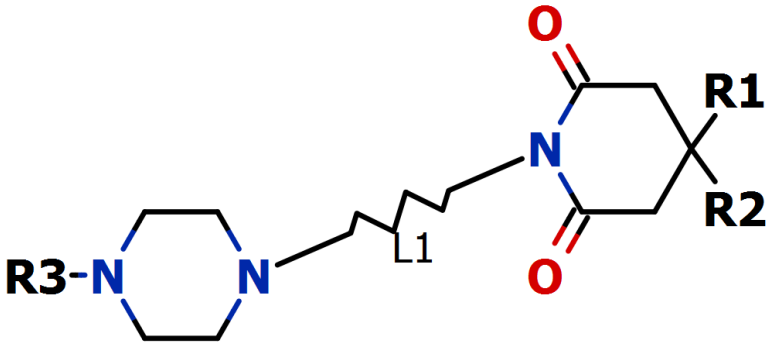
optibrium



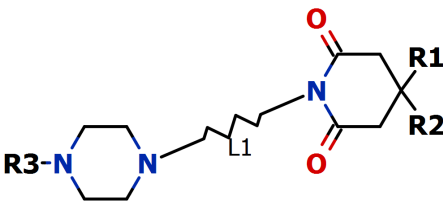
Focusing on the arylpiperazines

- One subset of these are Buspirone analogues

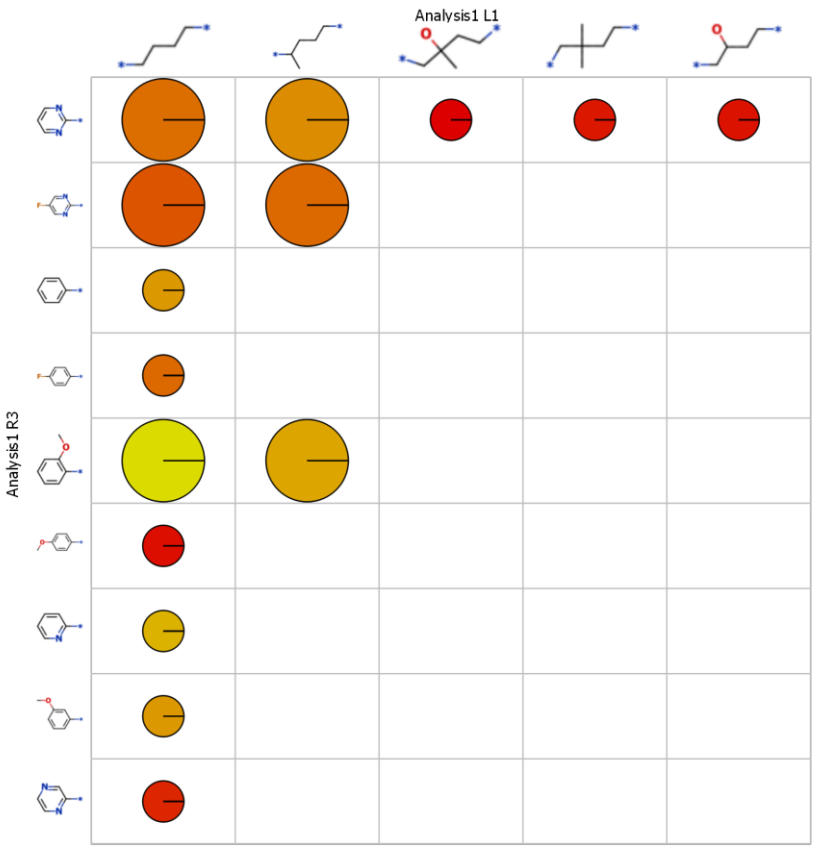
- Measured potency with pk_is between 5.8 and 8.7
- Measure stability (CYP3A4 half-life) between 3 and 80 minutes
- 20 analogues



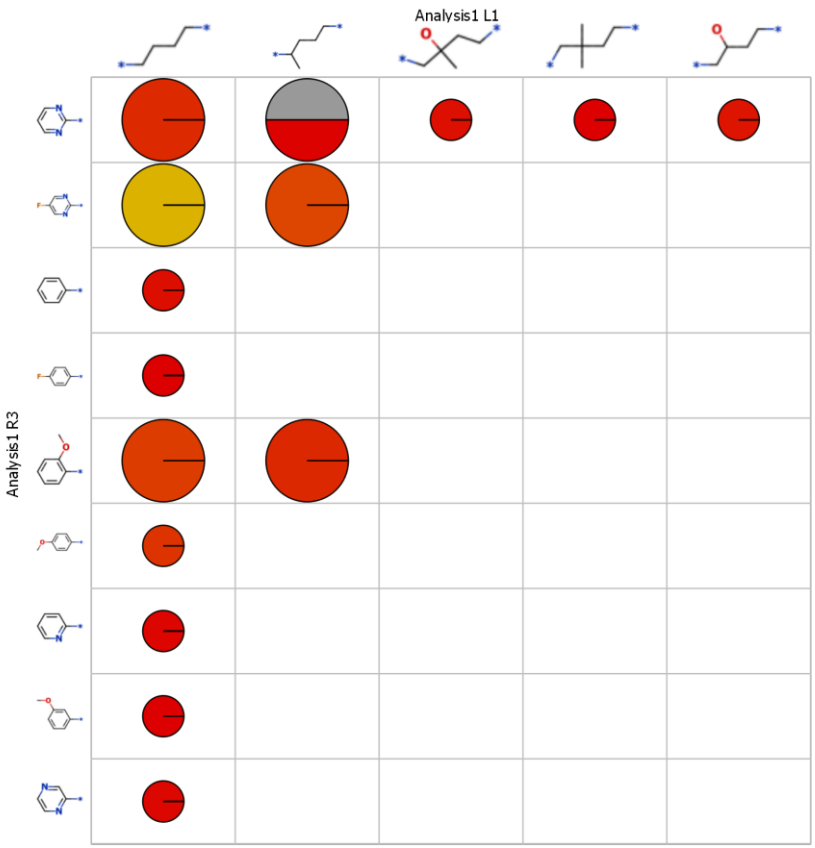
SAR tables



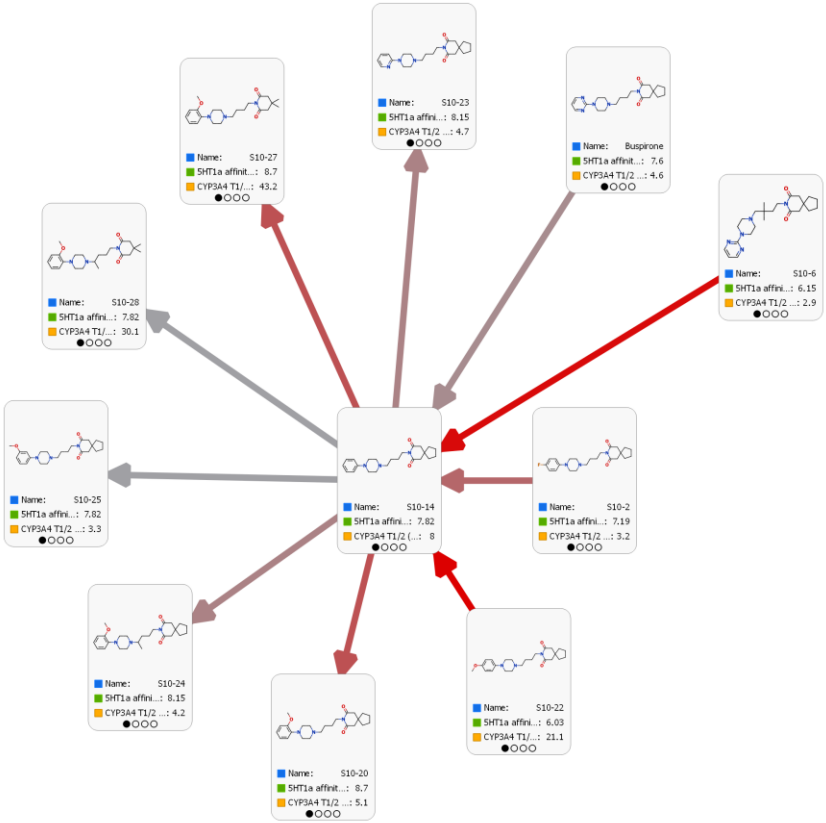
Potency



Stability

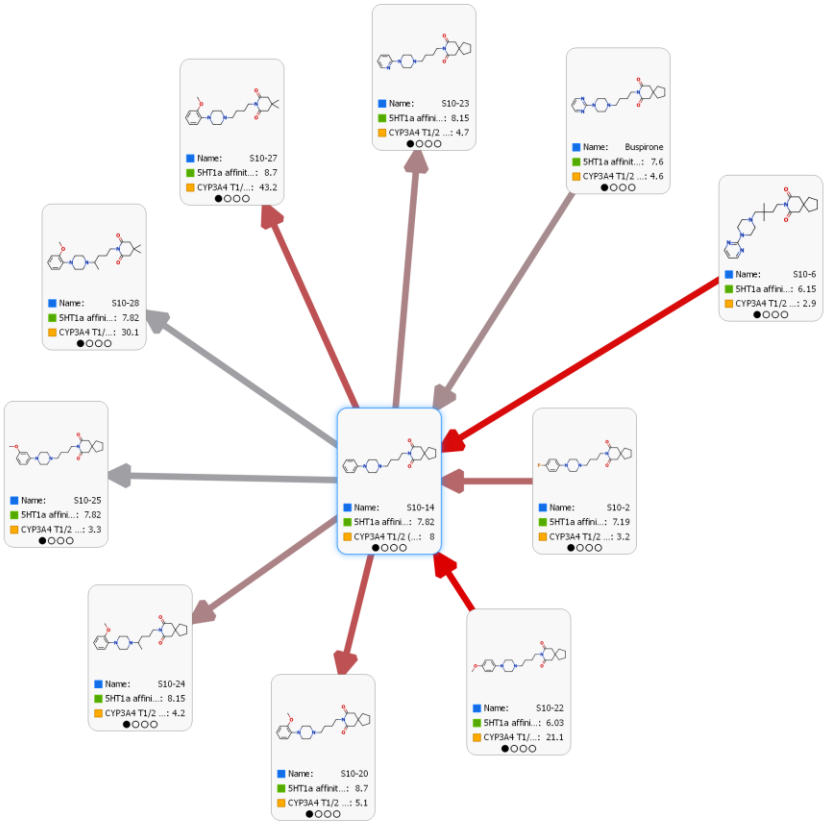


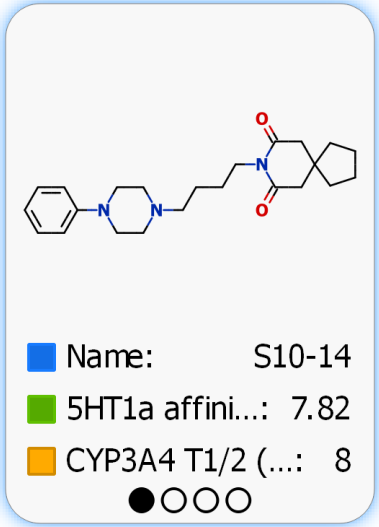
Activity neighbourhood



Activity neighbourhood

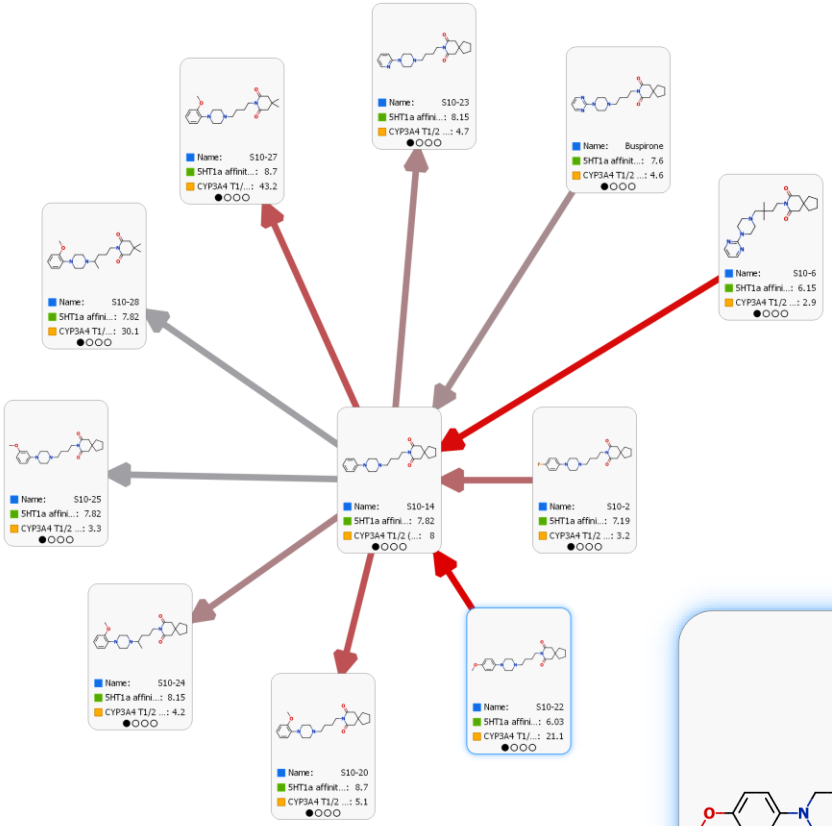
Good potency, poor stability

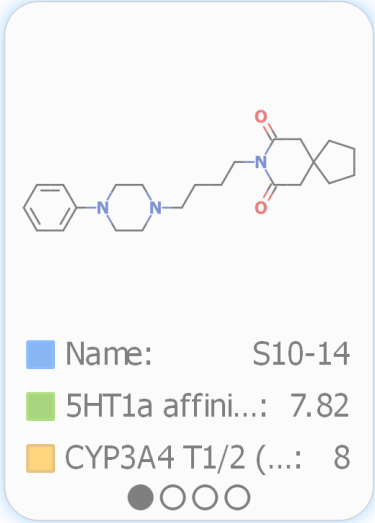


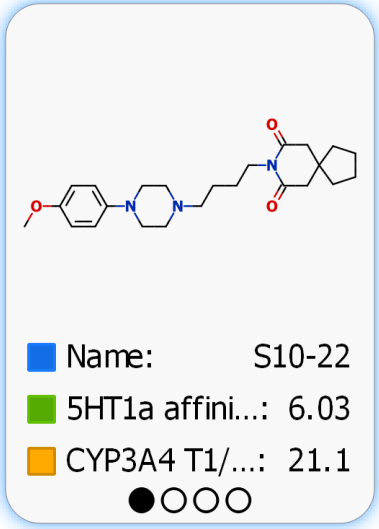


Activity neighbourhood

Good stability, poor potency

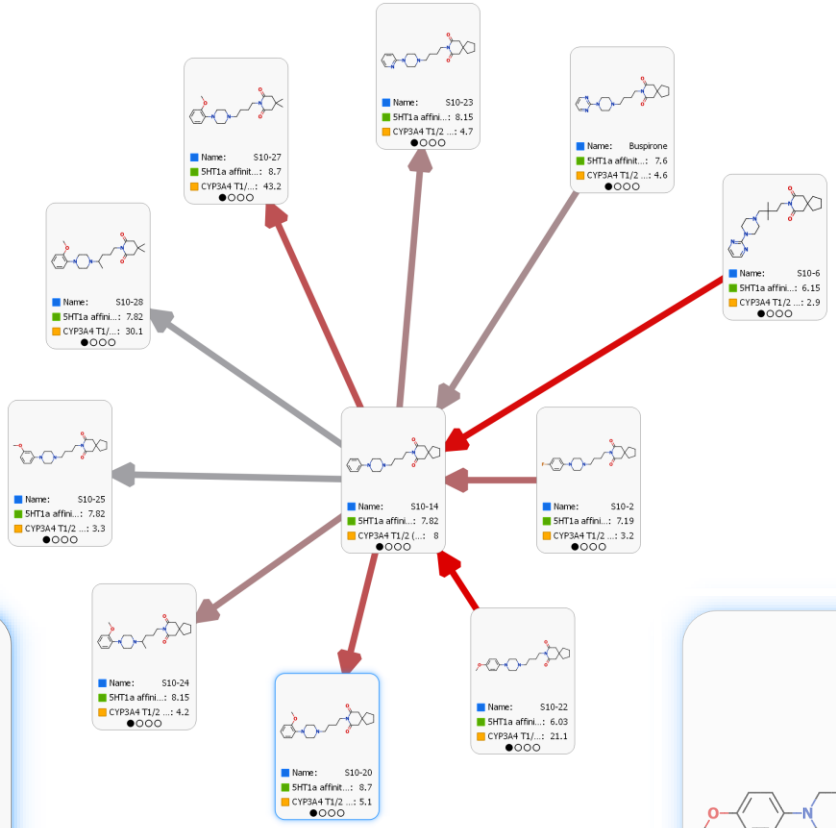


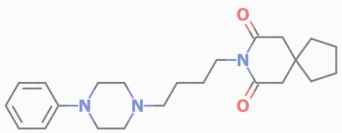




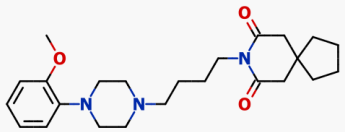
Activity neighbourhood

Good potency, poor stability

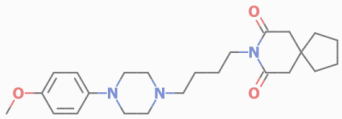




Name: S10-14
5HT1a affini...: 7.82
CYP3A4 T1/2 (...: 8



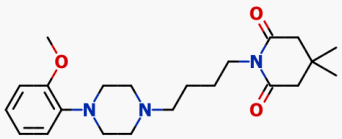
Name: S10-20
5HT1a affinit...: 8.7
CYP3A4 T1/2 ...: 5.1



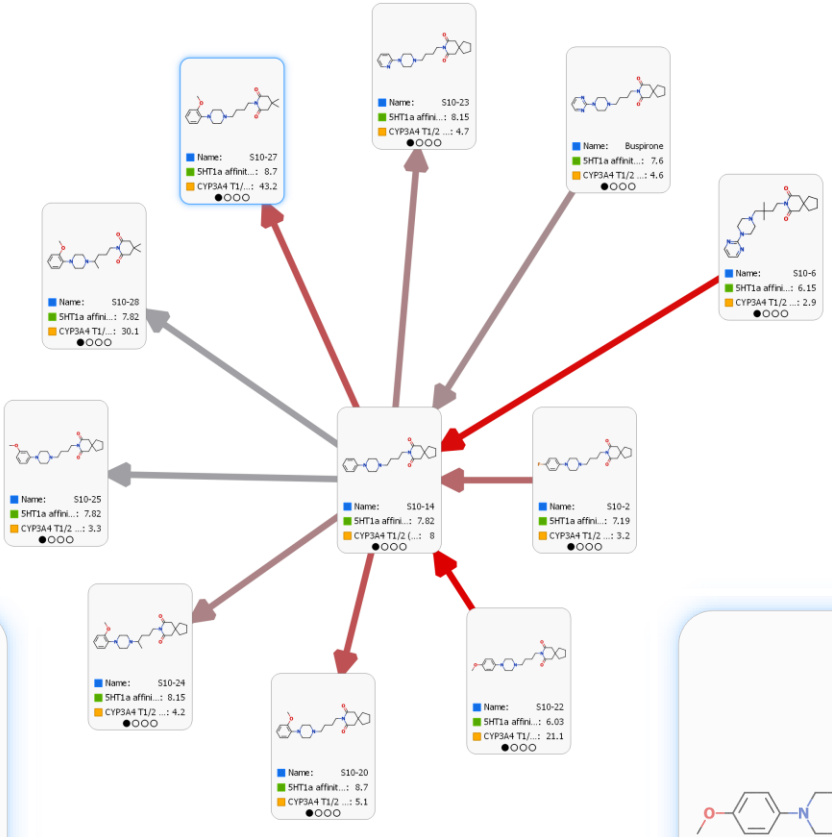
Name: S10-22
5HT1a affini...: 6.03
CYP3A4 T1/...: 21.1

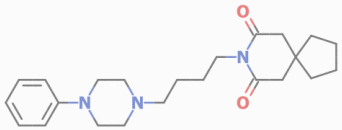
Activity neighbourhood

Best of both worlds

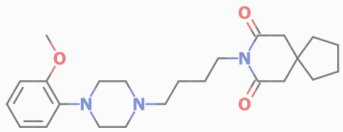


Name: S10-27
5HT1a affinit...: 8.7
CYP3A4 T1/...: 43.2

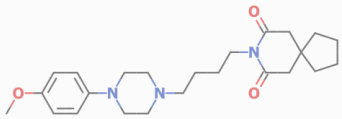




Name: S10-14
5HT1a affini...: 7.82
CYP3A4 T1/2 (...: 8



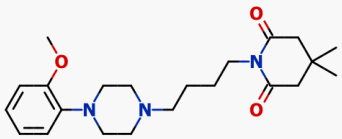
Name: S10-20
5HT1a affinit...: 8.7
CYP3A4 T1/2 ...: 5.1



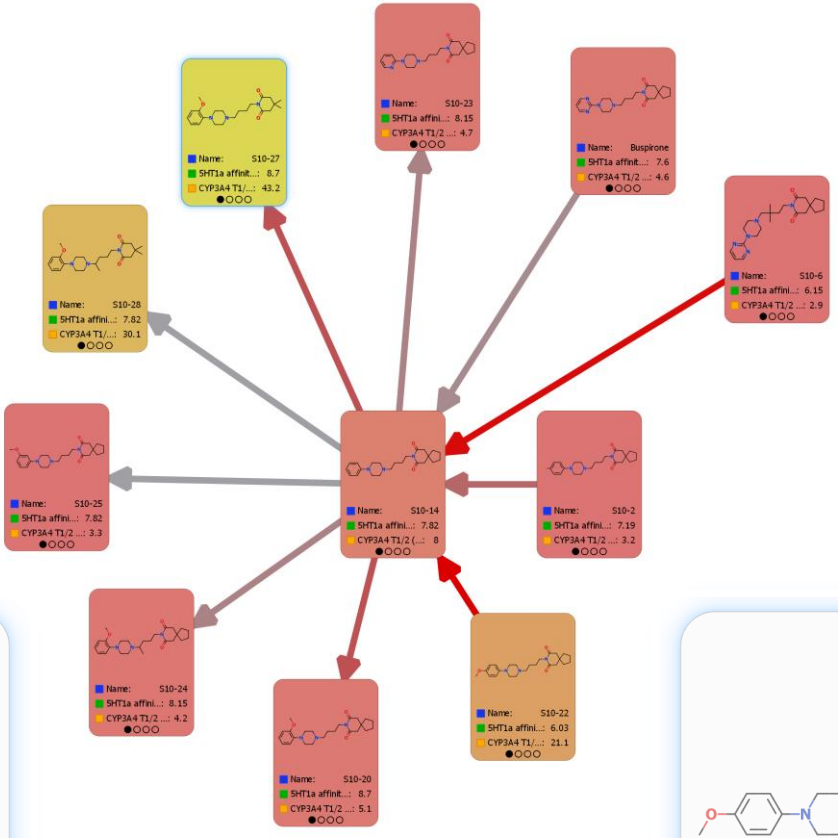
Name: S10-22
5HT1a affini...: 6.03
CYP3A4 T1/...: 21.1

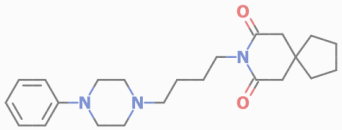
Activity neighbourhood

Best of both worlds

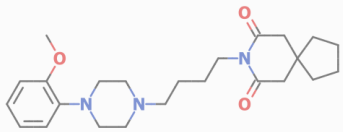


Name: S10-27
5HT1a affinit...: 8.7
CYP3A4 T1/...: 43.2

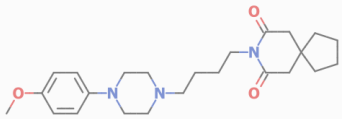




Name: S10-14
5HT1a affini...: 7.82
CYP3A4 T1/2 (...: 8



Name: S10-20
5HT1a affinit...: 8.7
CYP3A4 T1/2 ...: 5.1



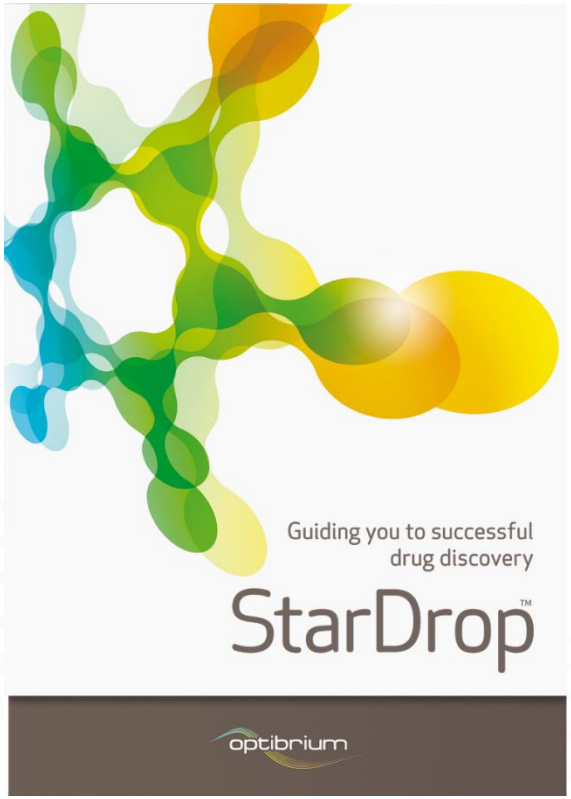
Name: S10-22
5HT1a affini...: 6.03
CYP3A4 T1/...: 21.1

Conclusions

- It is very often the network of inter-relationships between compounds that matter
  - These often influence the creation of the next compound
- The way we perceive our compounds depends upon those around it
  - Timing: have we explored the surrounding chemical space thoroughly enough to adequately evaluate a series?
  - Property data: do we have data of sufficient quality to confidently distinguish the good compounds?
- Visualisations that collapse or remove that network relationship and context always have the potential to bias our perception of the data.

Acknowledgements

- Matt Segall
- Peter Hunt
- Chris Leeding
- James Chisholm
- Hector Garcia Martinez
- Alex Elliott
- Sam Dowling
- Nick Foster



- Exhibition booth #417/516
- CINF 162: Modeling ABC transporters as potential DILI targets – Matt Segall
  - Weds 19 August 15:50 – 16:10 : Room 103