



Cheminformatics from the end-user perspective: Past, present and future.

Paul Greenspan,
Senior director of oncology chemistry, Takeda

April 11, 2016

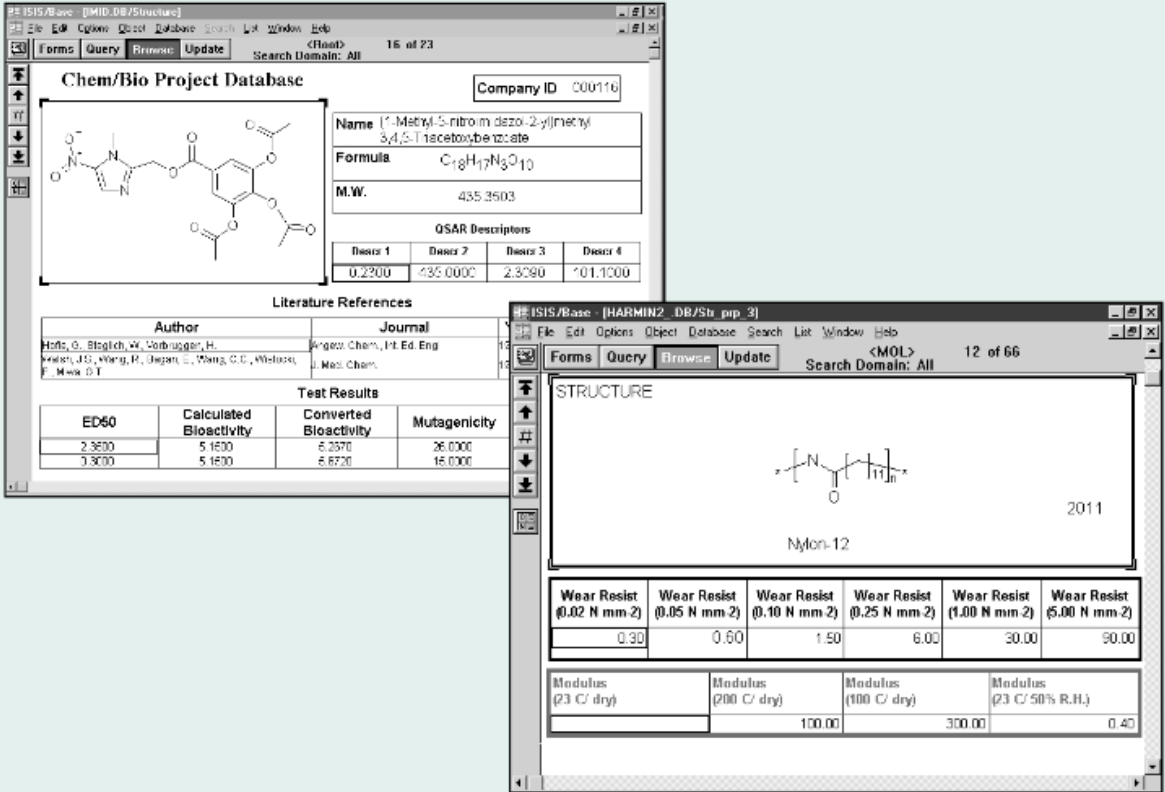
Thank you in advance for your patience!

- This will NOT be a technical presentation. Sorry about that!
- I have been an industrial medicinal chemist for 25 years.
 - 11 years at Novartis (arthritis, inflammation)
 - 14 years at Millennium/Takeda (oncology)
- I am definitely NOT a cheminformaticist or computational chemist, but I have a lot of interest in the field, and greatly appreciate the value.
- Today, I will present my perspective on the evolution of cheminformatics over the course of my career, and what key challenges lie ahead.

What were things like 25 years ago for a medicinal chemist?

- Typical chemistry throughput might be 10 compounds/chemist/month
- An “HTS” might be 10,000 compounds/month
- Very limited use of assays beyond primary screens.
 - 1 or 2 datapoints per compound.
- What was the state of “cheminformatics” 25 years ago?
 - Medicinal chemistry databases were just being introduced
 - MDL was the only game in town
 - Most project teams kept assay data in private databases (or spreadsheets)
 - Until ~2000, the key challenge was getting data into a searchable database

Remember when this was state of the art?



Chem/Bio Project Database

Company ID: 000116

Name: (7-Methyl-5-nitroimidazo[2-y]lmethyl 3,4,5-triacetoxibenzoate

Formula: $C_{18}H_{17}N_5O_{10}$

M.W.: 455.3503

QSAR Descriptors

Descr 1	Descr 2	Descr 3	Descr 4
0.2300	-0.350000	2.3090	-0.11000

Literature References

Author	Journal
Hofz, G., Stoglich, W., von Brugger, H.	Angew. Chem., Int. Ed. Eng.
Kaiser, J.S., Wang, R., Ujapan, E., Wang, C.C., Wilness, F., Meek, D.T.	J. Med. Chem.

Test Results

ED50	Calculated Bioactivity	Converted Bioactivity	Mutagenicity
2.3600	5.1600	6.3370	26.0000
3.3000	5.1600	6.6720	16.0000

ISIS/Base [HARMIN2_DB/Stu pup.3]

STRUCTURE

Nylon-12

Wear Resist (0.02 N mm ²)	Wear Resist (0.05 N mm ²)	Wear Resist (0.10 N mm ²)	Wear Resist (0.25 N mm ²)	Wear Resist (1.00 N mm ²)	Wear Resist (5.00 N mm ²)
0.30	0.60	1.50	6.00	30.00	90.00

Modulus (23 C° dry)	Modulus (200 C° dry)	Modulus (100 C° dry)	Modulus (23 C° 50% R.H.)
	100.00	300.00	0.40

- Customizable GUI, multiple display options, structure and data searching

Volume of data has exploded in past 25 years

More compounds

- Routine HTS screens of $>10^6$ compounds
- High-throughput synthetic chemistry
- New ultra-high-throughput screening approaches (eg. DNA-encoded libraries)
- Enormous “virtual” compound libraries.
- External vendors with vast catalogs of compounds

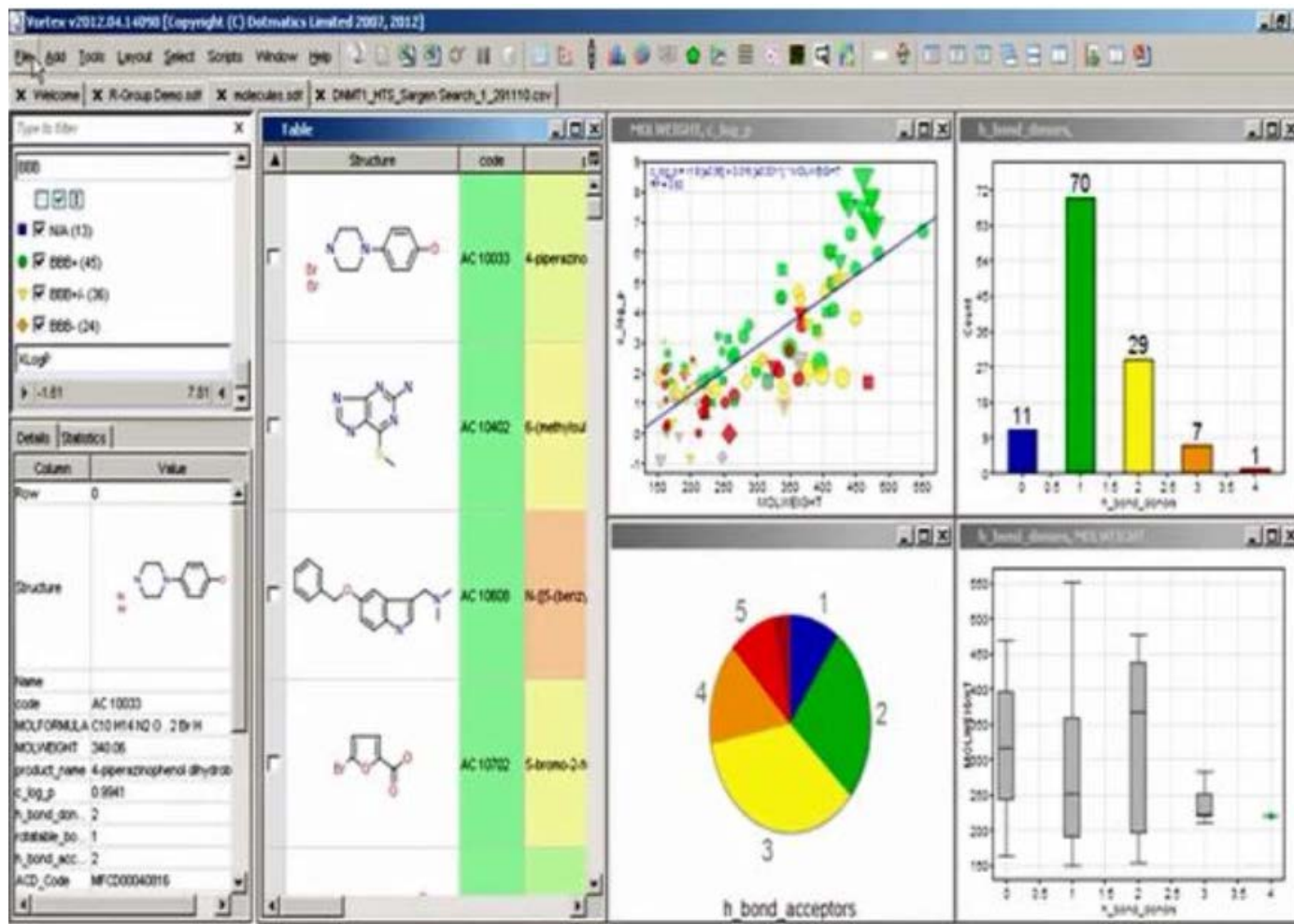
More data per compound

- Extensive cross target selectivity screening
- Broad target-class screens (eg. Kinome panels)
- Routine HT predictive ADMET screening
- Predictive modeling generating lots of “virtual” data
- Large external chem/biology databases (pubchem, chembl, etc.)

Cheminformatics has come a long way...

- Global, user-friendly chemistry/biology databases are commonplace (if not universal)
- Predictive modeling has become much more mainstream
- Broad implementation of electronic notebooks has made even “raw data” accessible.
- Entirely new ways of analyzing data have taken hold:
 - Dynamic querying and visualization tools (spotfire, etc.)
 - Multi-parameter optimization methodologies allow more “holistic” analysis
 - Specialty tools (MMP, activity landscape analysis, etc.)
 - Clustering, framework analysis

Chemistry Dashboards integrate data seamlessly



Example: Dotmatics Vortex

These changes have redefined the challenge in fundamental ways

- 25 years ago, the goal was to make data available to allow chemists to review SAR data manually.
 - We couldn't envision tools to allow for more than that.
 - The datasets were small and simple enough to make this practical
- Today, datasets are far too large and complex for chemists to consume, analyze and draw conclusions manually from the data they receive.
- The key cheminformatics challenge is to enable chemists to make optimal use of all this data:
 - Construct testable hypotheses
 - Effectively prioritize design ideas
 - Assist chemists' imagination in generating new approaches

There are several challenges in supporting med chemists in working with large datasets:

- Med chemists don't like math!
 - We tend to think visually, rather than mathematically.
 - Outcomes of statistical analyses must be conceptually straightforward.
- Chemists don't deal well with uncertainty:
 - A chemical structure is absolute. Biological data is not.
- There is no perfect way to parameterize a chemical structure:
 - Chemists may not agree with calculated similarities, clustering, etc.
 - Meaning of atom connectivities can be very context-dependent.

Dumbing down the data

- If chemists don't like math, and struggle to conceptualize large datasets, then let's keep it simple.
- Create “rules” that any idiot can obey:
 - Lipinski Rule of 5.
 - Internal cut-offs imposed by many pharma organizations
- But can this possibly be right?
 - Aren't these things context dependent?
 - Is MW of 495 really infinitely better than MW of 505?
 - If lipophilicity is low, couldn't we back off on our MW cut-off?

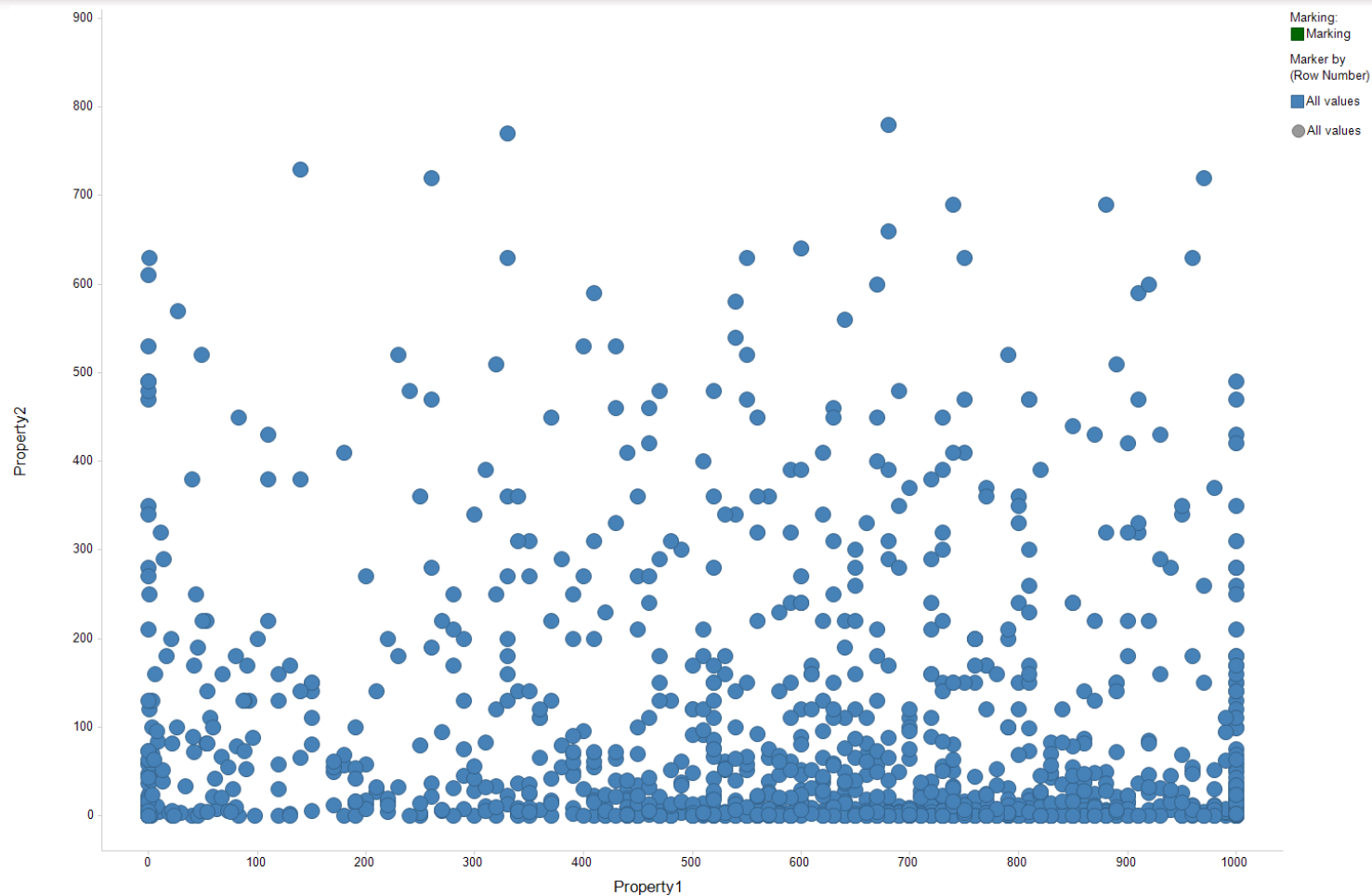
Is there a better approach?

- Unintuitive mathematical constructs have limited appeal.
- Oversimplification can lead to erroneous decision-making
- Datasets are too large and complex to expect a chemist to retrieve all potential value through manual inspection.
- How do we help chemists in a way that plays to their strengths?
 - Data visualization
 - Computational identification of data “gems”

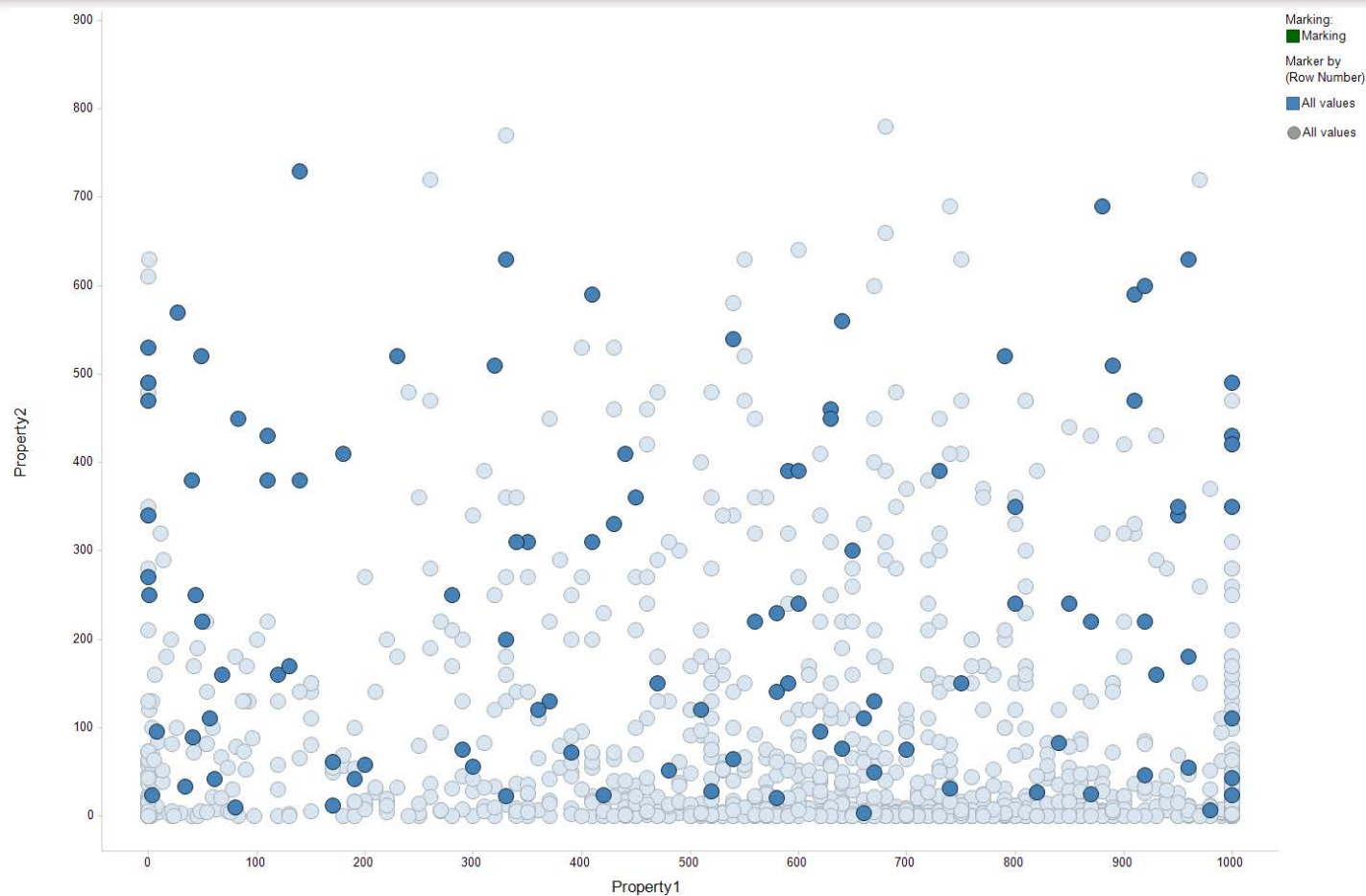
Visualization: a big breakthrough

- Spotfire introduced the concept of interactive visualization to medicinal chemistry and drug discovery
 - Bridged the gap between manual SAR analysis and statistical methods.
 - Allowed chemists to be in control: view data from variety of perspectives, pose questions that can only be answered with aggregate data.
 - Outputs are visual, not mathematical.
 - Allowed for real-time, iterative data interrogation and hypothesis generation

Example: No obvious trends across data-set

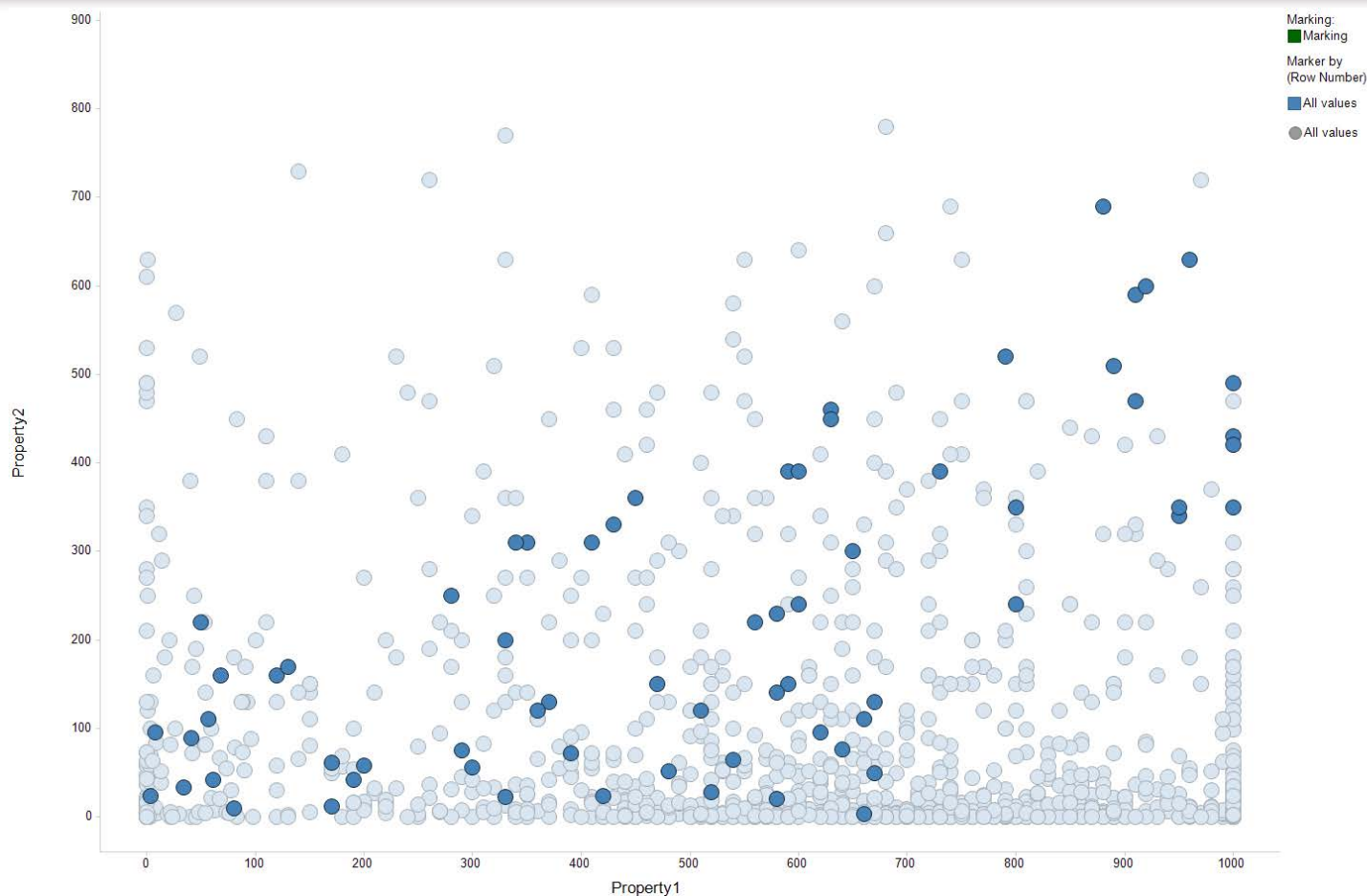


Is there a trend if we only look at amines?



- chemistry queries with visual output

How about amines with $\log p < 3$?

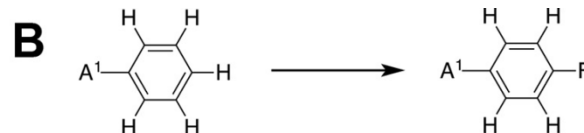
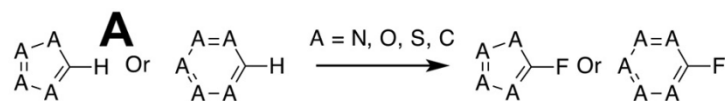


- Explore additional data relationships interactively
- Create testable hypotheses

Another Breakthrough: finding the data “gems”

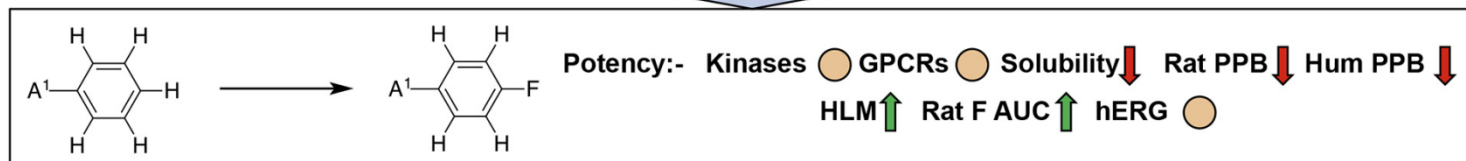
- Sometimes, the most important data is “small”:
 - The comparison of a few datapoints may tell a critical story
- But how do chemists pick that out from all the noise?
- Cheminformatics has helped chemists to home in on key data:
 - Matched molecular pair analysis
 - Activity landscapes

The Power of Matched Molecular Pair Analysis



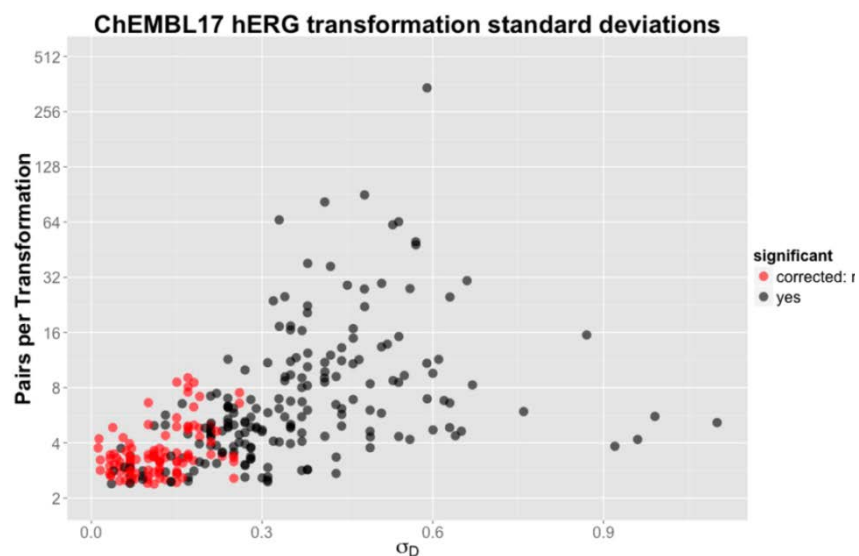
Assay Endpoint	n	Δ Mean	SE	SD	% Improve	Source
log $D_{7.4}$ (all H to F)	9902				7.7 (2 fold)	Papadatos [20]
log Aq. Solubility	1572	-0.10	<0.01	0.34	24 (1.6 fold)	Gleeson [19]
log Aq. Solubility (all H to F)	4273				17 (2 fold)	Papadatos [20]
Cytochrome P450 inhib pIC_{50}						
1A2	1244	-0.03	<0.01	0.45	32 (1.6 fold)	Gleeson [19]
2C9	2683	+0.04	<0.01	0.37	23	
2C19	1860	+0.04	<0.01	0.37	22	
2D6	2087	+0.03	<0.01	0.41	23	
3A4	2297	+0.07	<0.01	0.42	21	
Permeability (log nM/s)	2848	+0.01	<0.01	0.32	15 (2 fold)	Gleeson [19]
Rat in-Vivo Unbound Clearance	96	0.11	0.04			Sutherland [21]
Potency at target classes pIC_{50} (all H to F)						Hajduk [8]
Kinases (7)	942				3.8 (10 fold)	
Class 1 GPCR (9)	642				1.4 (10 fold)	
Others (14)	1003				3.7 (10 fold)	
hERG pIC_{50}	1572	-0.10	0.01	0.34	24 (2 fold)	Gleeson [19]
hERG pIC_{50} (all H to F)	4243				18 (2 fold)	Papadatos [20]

Assay Endpoint	n	Δ Mean	SE	SD	% Improve	Source
log $D_{7.4}$	252	+0.18		0.44	4.9 (3.2 fold)	Dossetter [15]
log Aq. Solubility (from solid)	711	-0.22	0.02	0.36	34 (all >0)	Leach [12]
Human PPB K_i	171	+0.06	0.02	0.29	65 (all >0)	Leach [12]
Rat PPB K_i	407	+0.15	0.01	0.29	77 (all >0)	Leach [12]
AZ HLM log Cl_{int}	497	-0.06	0.02	0.36	8.6 (3.2 fold)	Dossetter [15]
Pf HLM log Cl_{int}	491				9.2 (2 fold)	Lewis [14]
Rat Oral Bio-availability log AUC	551	+0.09	0.03	0.65	55 (all >0)	Leach [12]
Potency at target classes pIC_{50}						Swiss bio-isoeter – ChEMBL data-mining [11]
Kinases (57)	291	-0.06	0.03	0.52	28 (2 fold)	
Class 1 GPCR (110)	1305	+0.00	0.02	0.37	27 (2 fold)	
Ion Channel VGC (9)	34	+0.02	0.06	0.38	18 (2 fold)	
hERG	8	+0.04	0.09	0.24	13 (2 fold)	

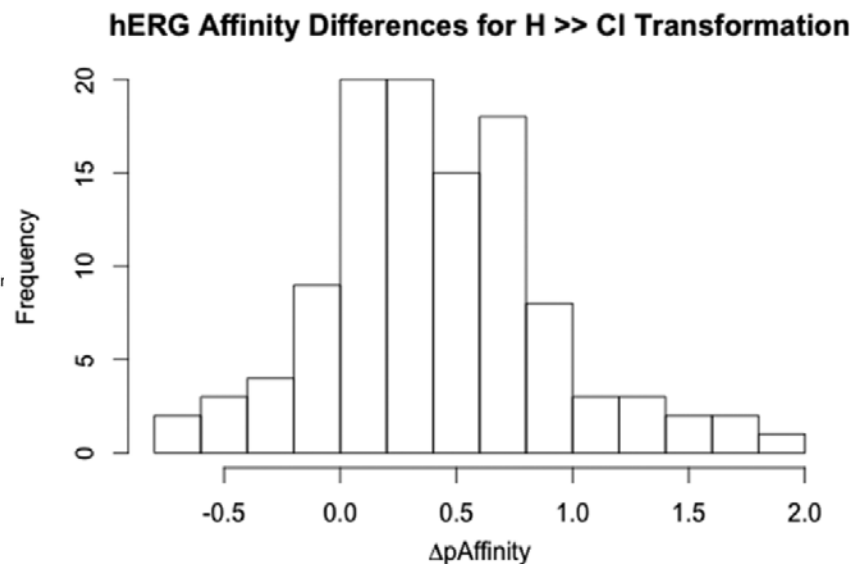


Dossetter, et. al. Drug Discovery Today, Vol. 18, p. 724

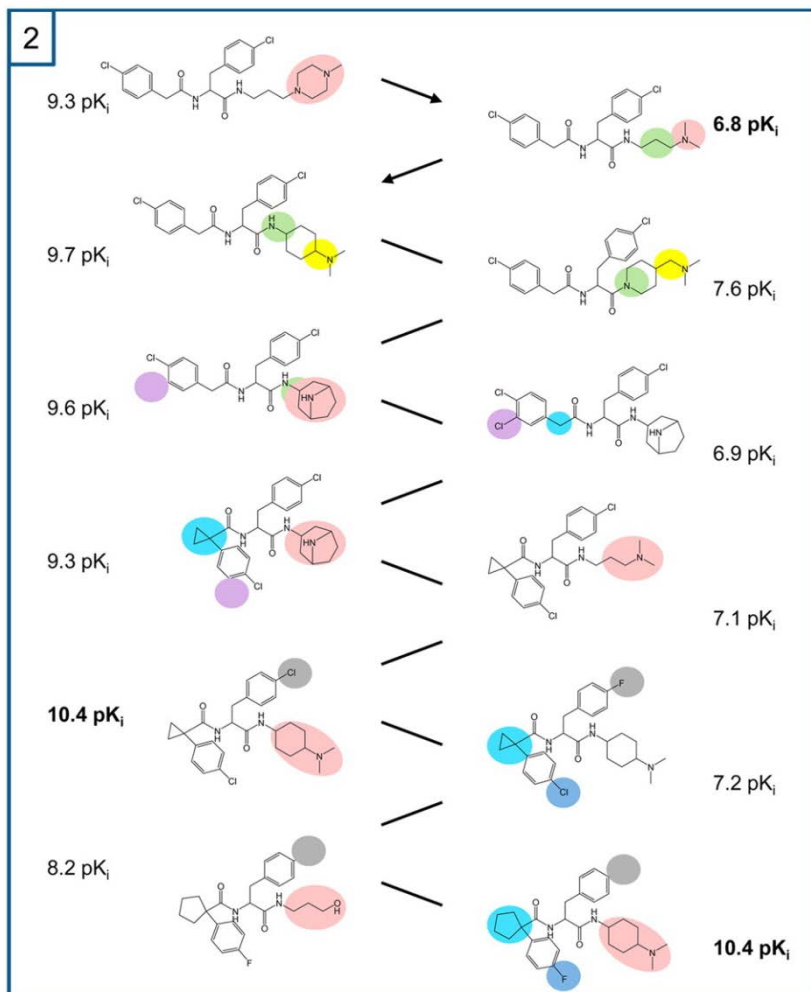
Need for Enhanced rigor with MMPA?



Kramer et. al. J. Med. Chem. 2014, 57, 3786



- 4 pairs sufficient to identify significant differences with homogenous data.
- 10-20 pairs needed if data comes from different assays.



- Vasopressin V1a data from ChEMBL
- Analysis capture key SAR inflection points
- Pulling this data manually out of a large database would be difficult or impossible.

Paradigm shift: Multi-parameter optimization

- Historically, chemists have relied on filters for decision-making
 - Selection of compounds for secondary, tertiary screening
 - Choosing compounds to synthesize or purchase.
- Very simple to implement and conceptualize
- Serious drawbacks:
 - Greatly exaggerates small differences in parameter values
 - Overly rigid: filter values not impacted by other parameters
 - Order of filters can have unintended consequences:
 - Good compound can be lost early if it barely misses the first filter.
- MPO allows chemist to take all parameters into account simultaneously

Marriage of visualization and MPO: Golden Triangle

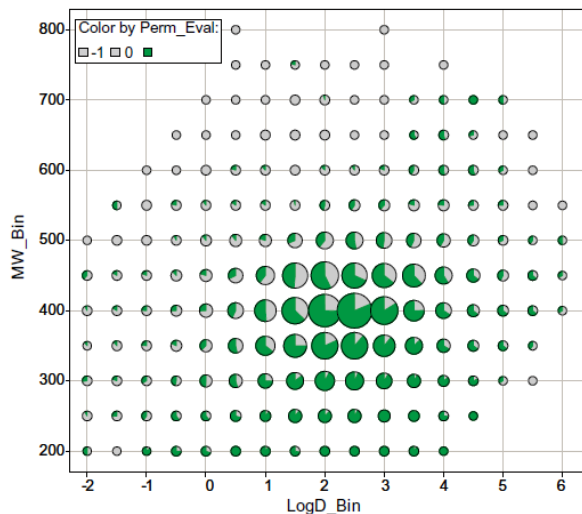


Figure 2. Comprehensive in vitro Caco-2 AB permeability trends across molecular weight and log D. 16,227 total records sized by record count.

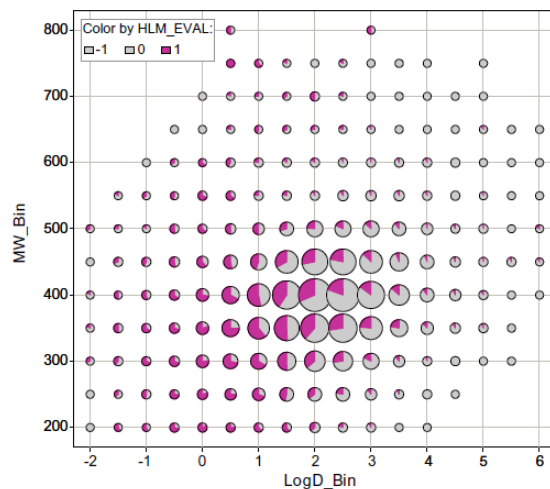


Figure 3. Comprehensive in vitro HLM clearance trends across molecular weight and log D. 47,018 total records sized by record count.

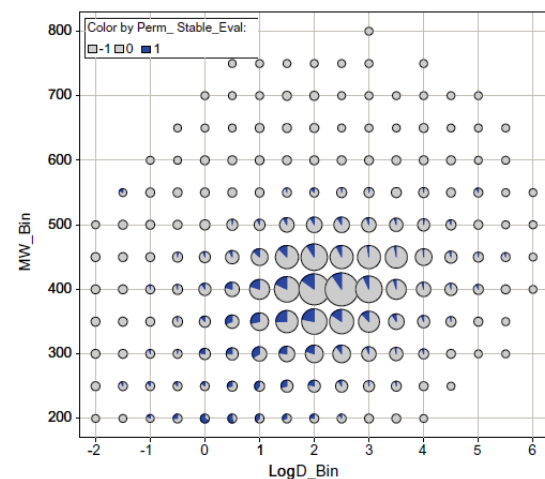
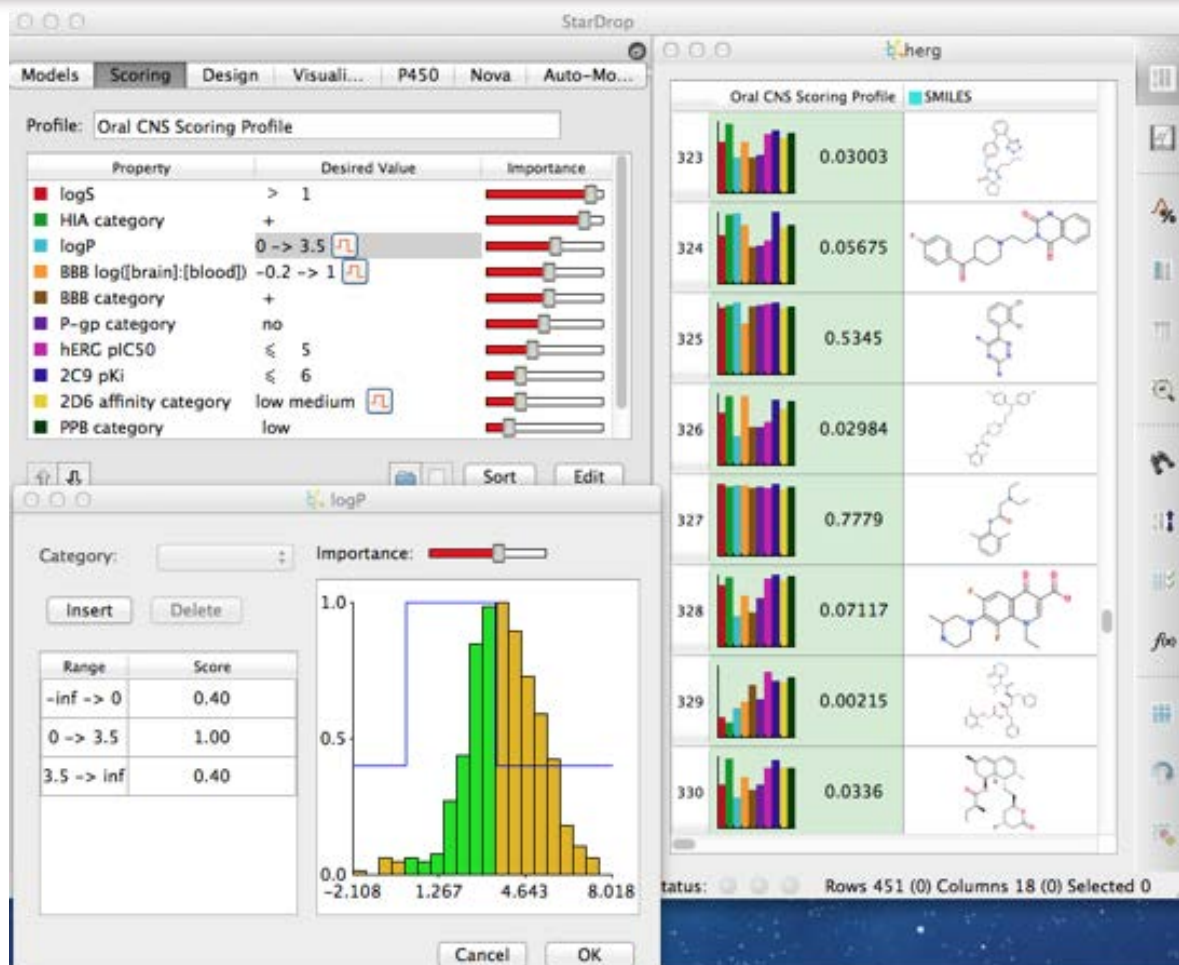


Figure 4. Comprehensive combined in vitro clearance and in vitro permeability trends across molecular weight and log D. 16,090 total records sized by record count.

T. W. Johnson et al.. Bioorg. Med. Chem. Lett. 19 (2009) 5560–5564

- Attempt develop more robust model for PK optimization
- Case is made primarily through visualization of multi-dimensional data

Probabilistic Scoring in Stardrop



- Stardrop allows chemist to control parameter weighting and selection
- Visualization allows chemist to readily see impact of each parameter

Predictive modeling: then and now

- Pitfalls of predictive modeling in the 90's:
 - Focus on building “global” models that try to explain everything.
 - Use of “opaque” statistical methods (PLS, PCA)
 - Lack of clarity regarding limits in predictiveness
- Predictive modeling fell out of favor:
 - Frustration of chemists who didn't understand models, and couldn't determine their limitations.
 - Backlash from “overhype” (companies overselling modeling software)
 - No good way to incorporate into chemistry workflow
- We are now seeing a resurgence in predictive modelling:
 - Better understanding of limitations and appropriate uses.
 - Greater focus on local models.
 - Visualization tools allow chemists to interact with models, and understand drivers of predictions

What has this innovation given us?

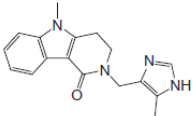
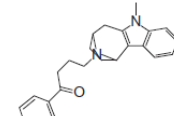
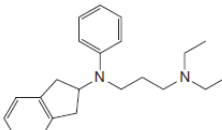
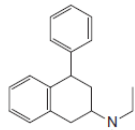
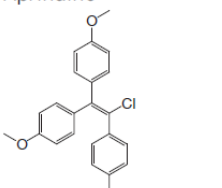
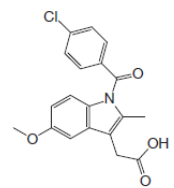
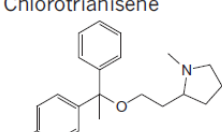
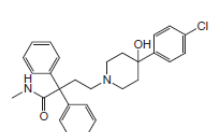
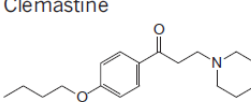
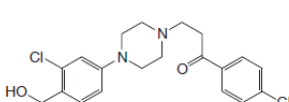
- Chemists can now effectively interrogate large datasets, discover trends, and form hypotheses.
- Chemists can find the “data gems” that could easily be lost in the noise of large data-sets.
- Chemists can apply predictive modeling to real-world problems, and understand when and how it can be used.
- Chemists can be much more sophisticated in prioritization and decision-making

So, what are the next challenges?

- Better utilization of external data:
 - Integration of large external databases with internal tools.
 - Effective means of handling heterogeneous data-sets.
 - “Real-time” data extraction and collation
- Better integration of bio-informatics and cheminformatics:
 - Improved methods for prediction of potential targets and off-targets.
 - target-hopping
 - phenotypic screening
- Better integration of informatics tools into chemistry workflows
- Help chemists manage their own pitfalls.

SEA: Predicting activity via chemical similarity

Table 1 | New drug-off-target predictions confirmed by *in vitro* experiment

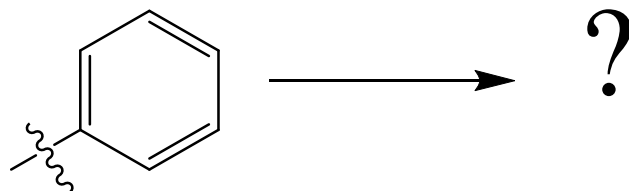
Drug	Closest chEMBL molecule	Tc value	Target	SEA <i>E</i> value	IC ₅₀ (μM)	Closest known target	BLAST <i>E</i> value
 Alosetron		0.25	HTR2B	10.6×10^{-17}	0.02	KCNH7	3.6×10^2
 Aprindine		0.38	HRH1	5.0×10^{-26}	0.78	SCN5A	3.3×10^{-1}
 Chlorotrianisene		0.31	COX-1	1.9×10^{-17}	0.16	ESR1	9.0×10^2
 Clemastine		0.31	SLC6A4	1.1×10^{-14}	0.42	KCNH2	6.1×10^1
 Dextropropriofen		0.36	DRD4	1.5×10^{-17}	4.1	SLC6A3	2.3×10^2

Lounkine, et. al Nature, vol. 486, p.361

- Predictions derived from analysis of ChemBL database
- Tremendous potential value for phenotypic screening

Computational approaches can help chemists to avoid pitfalls:

- Over-interpretation of statistically insignificant SAR
 - Too few datapoints, insignificant data differences.
 - Assist chemist to design experiments to enhance robustness.
- Tendency to form SAR assumptions, and not challenge them sufficiently.
 - “There’s no way an amine would be tolerated in that location...”
 - What is the basis of the assumption? Is it valid? How would it best be tested?
- SAR “white-space” exploration is not usually done systematically.



Thank you for your
attention!!

Enjoy the rest of the
symposium