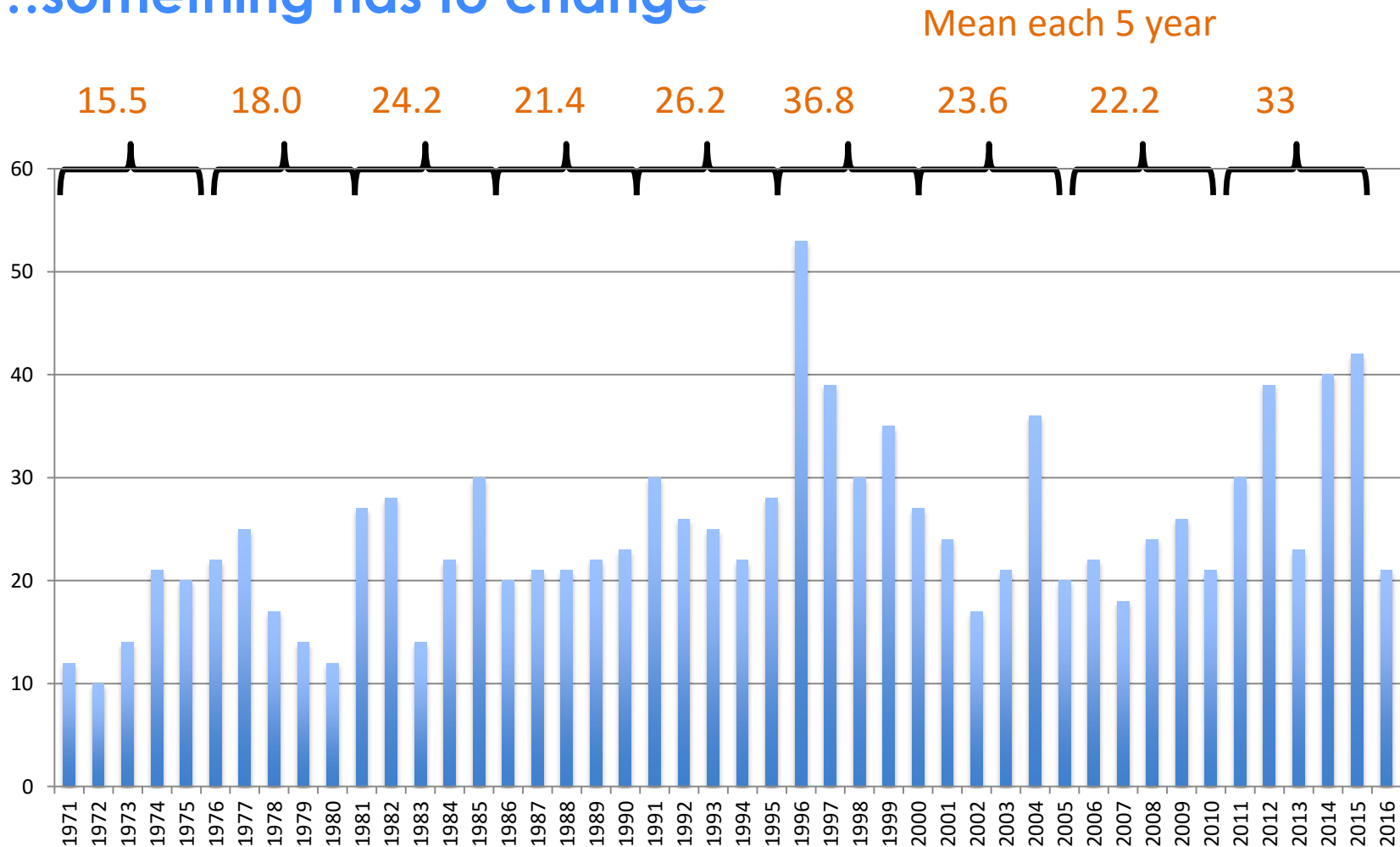


Extracting medicinal chemistry knowledge by a secured Matched Molecular Pair Analysis platform: standardization of SMIRKS enables knowledge exchange

Dr Alexander Dossetter
MedChemica

CICAG Structure Representaton meeting, 22 June 2017
Liverpool University, UK

NCE Drug Approval have not increased enough ...something has to change



Data - Federal Drug Administration Website <https://www.fda.gov>

Attrition in the Pharmaceutical Industry: Reasons, Implications, and Pathways Forward

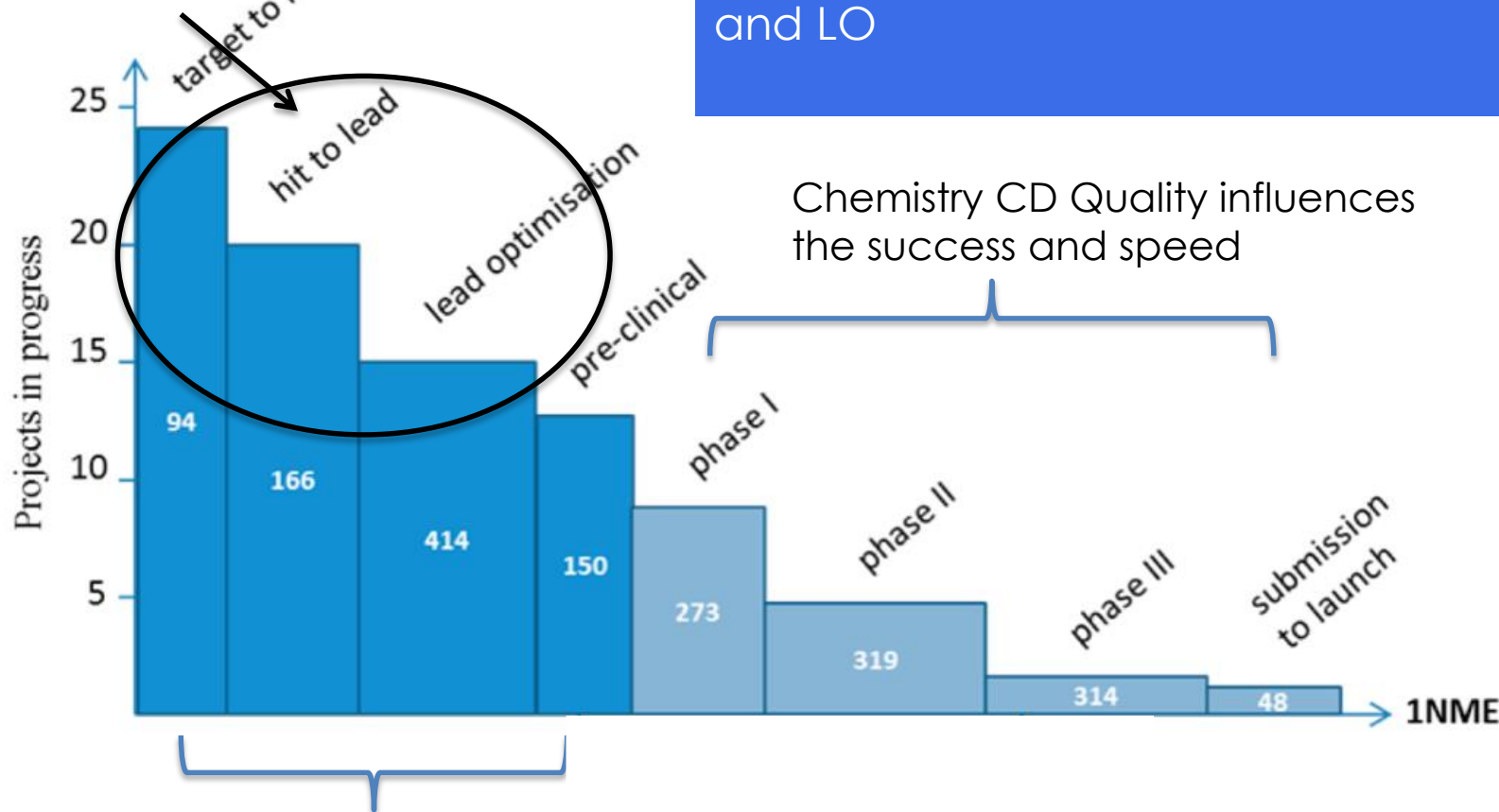
By Alexander Alex, C. John Harris, Dennis A. Smith; Wiley 2016



Actual spending / Chemistry everywhere

Better Knowledge =
Fewer Compounds =
Lower the cost

Medium sized companies R&D spend in one year \$1.7 billion 34% is spent in H2L and LO



Chemistry controls the productivity
and quality

Paul, S. M. *et al* How to improve R&D productivity: the pharmaceutical industry's grand challenge, *Nat. Rev. Drug Discovery* **2010**, 9, 203



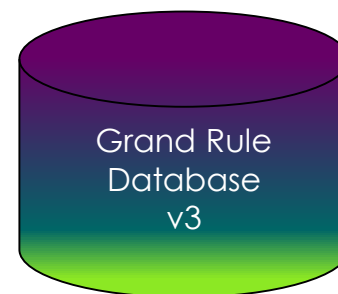
Genentech
A Member of the Roche Group

Where is the “Handbook of Medicinal Chemistry”?

- **Case study collections**
- **“War stories” & anecdotes**
- **Broad highly general rules (eg Lipinski)**

Where is the evidence based quantitative guide to medicinal chemistry?

Here is the story of the building of the ‘Grand Rule Database’.
A Multi-pharma Med Chem Textbook based on a thorough AI study.

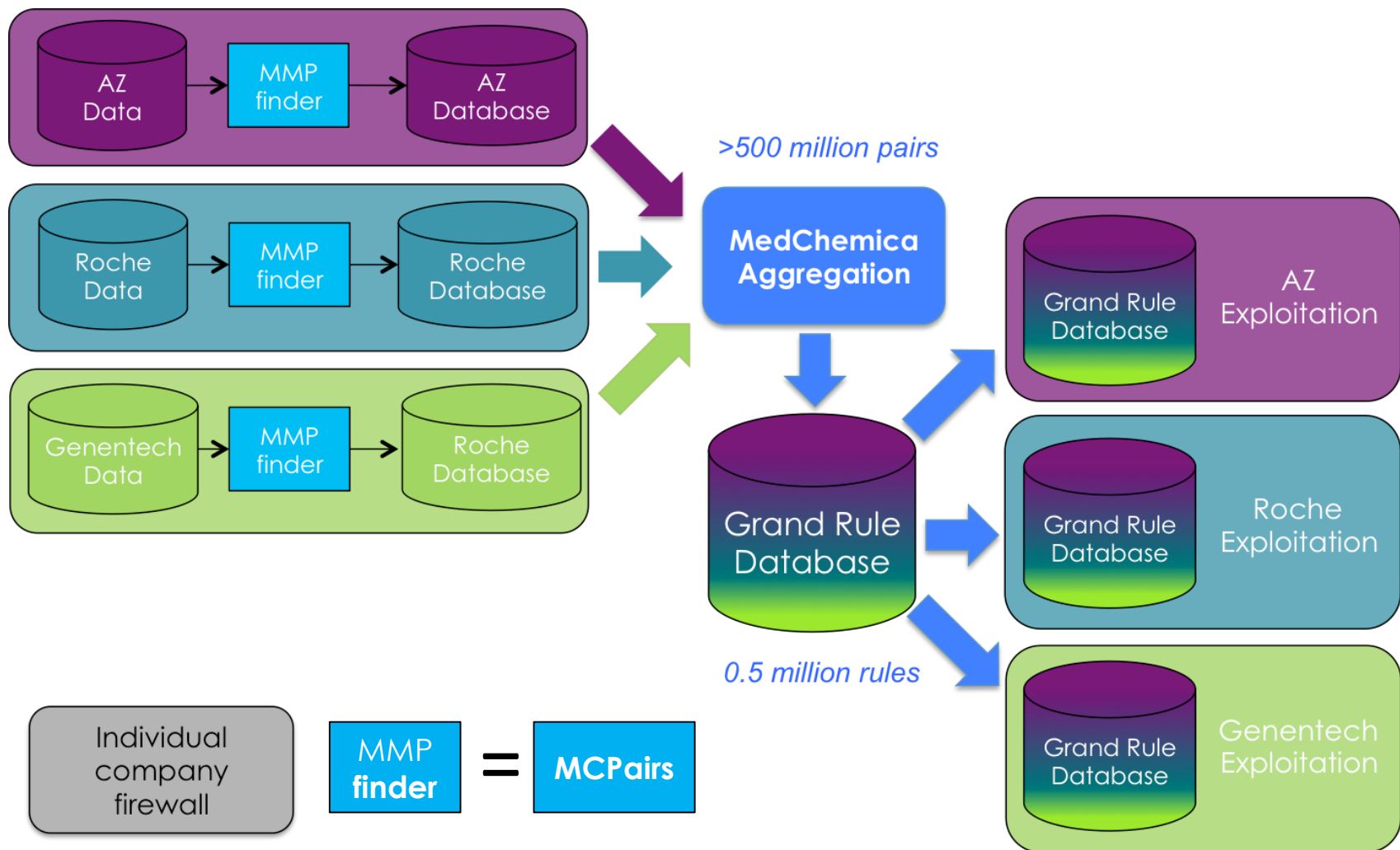


What we actually need is *UNSUPERVISED* Machine Learning

What?	Where?	Why?
Large datasets	Large Pharma	Access to all the "actives" and "in-actives"
Algorithms to extract structure	Matched Molecular Pair Analysis	All combinations considered [$O(n^2)$ problem], accurate structures, speed, finds counter intuitive Rules
Compute resource	Within Pharma	IP secure
Storage	Secure with VM	Multi-T-bytes
Ability to visualize and apply the results	Modern web tools and REST API	Chemists understand it and use it

Grand Rule database

Better medicinal chemistry by sharing knowledge not data & structures

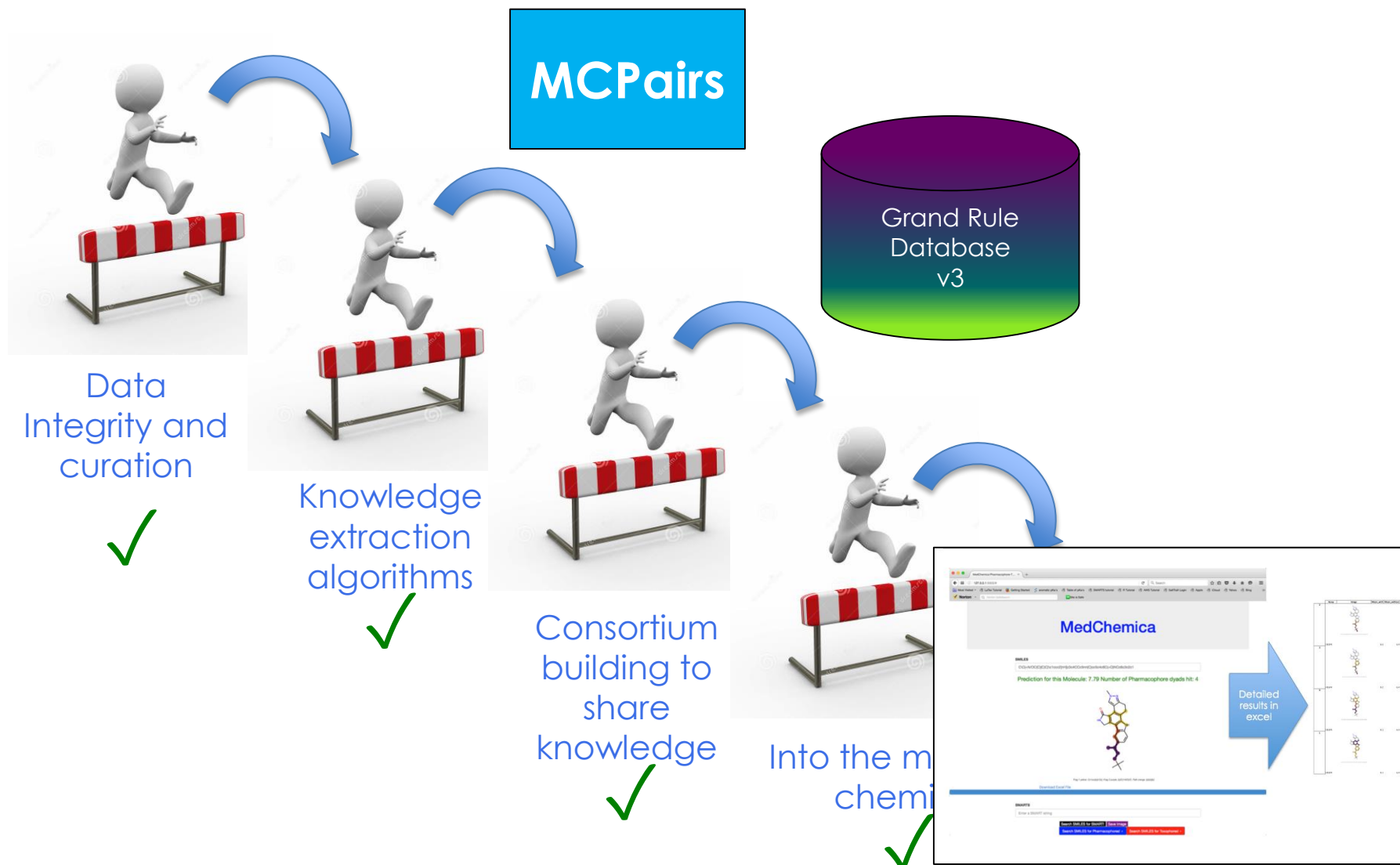


Kramer, C.; Ting, A.; Zheng, H.; Hert, J.; Schindler, T.; Stahl, M.; Robb, G.; Crawford, J.; Blaney, J.; Montague, S.; Leach, A. G.; Dossetter, A. G.; Griffen, E. J. *Learning Medicinal Chemistry ADMET rules from Cross-company MMPA* J.Med.Chem. Submitted.

Finding Matched Pairs and Chem-informatics

- Challenge:
 - Matched Pair finding is an $O(n^2)$ process so will be “BigData”
 - What is the best matched pair finding technique?
 - Once the pairs are found, how do you encode the output so knowledge can be shared securely?
 - Once there is knowledge how do chemists use it?

Barriers Broken to Sharing Knowledge



Standardisation (Units, Species, Routes, Aggregation)

11.1.2 About standardisation of Stereochemistry

There are several different levels of stereochemical definition that a compound can have. These and their required treatment are detailed below:

New – stereochemistry standard

22 standard units
Linear scale / categorical

id	description	long hand units	unit	related bao endpo
1	from example MCDK or p-gp efflux ratio, LogD(pH7.4)	unitless	scalar	BAO_0002129
2	from aqueous solubility assays - special note log(Molar) scalar	logarithm base 10 (Molar concentration)	log10(M)	BAO_0002135
3	from PAMPA			
4	from albumin b			
5	from permeabi			
6	from hepatic c			
7	from microsom			
8	from whole org			
9	from in-vivo O mum concentra			
10	from in-vivo l pound half life			
11	from in-vivo b			

Table 3. Assay standards

Biological, physical chemistry or toxicological endpoint	Species or isoform	Assay standard(s)
Plasma protein binding	Rat	Tolbutamide, Warfarin
	Mouse	Tolbutamide or Quinidine
	Dog	Tolbutamide, Quinidine or Warfarin
In vitro hepato		

Shared Assay Standard
Linear scale / categorical

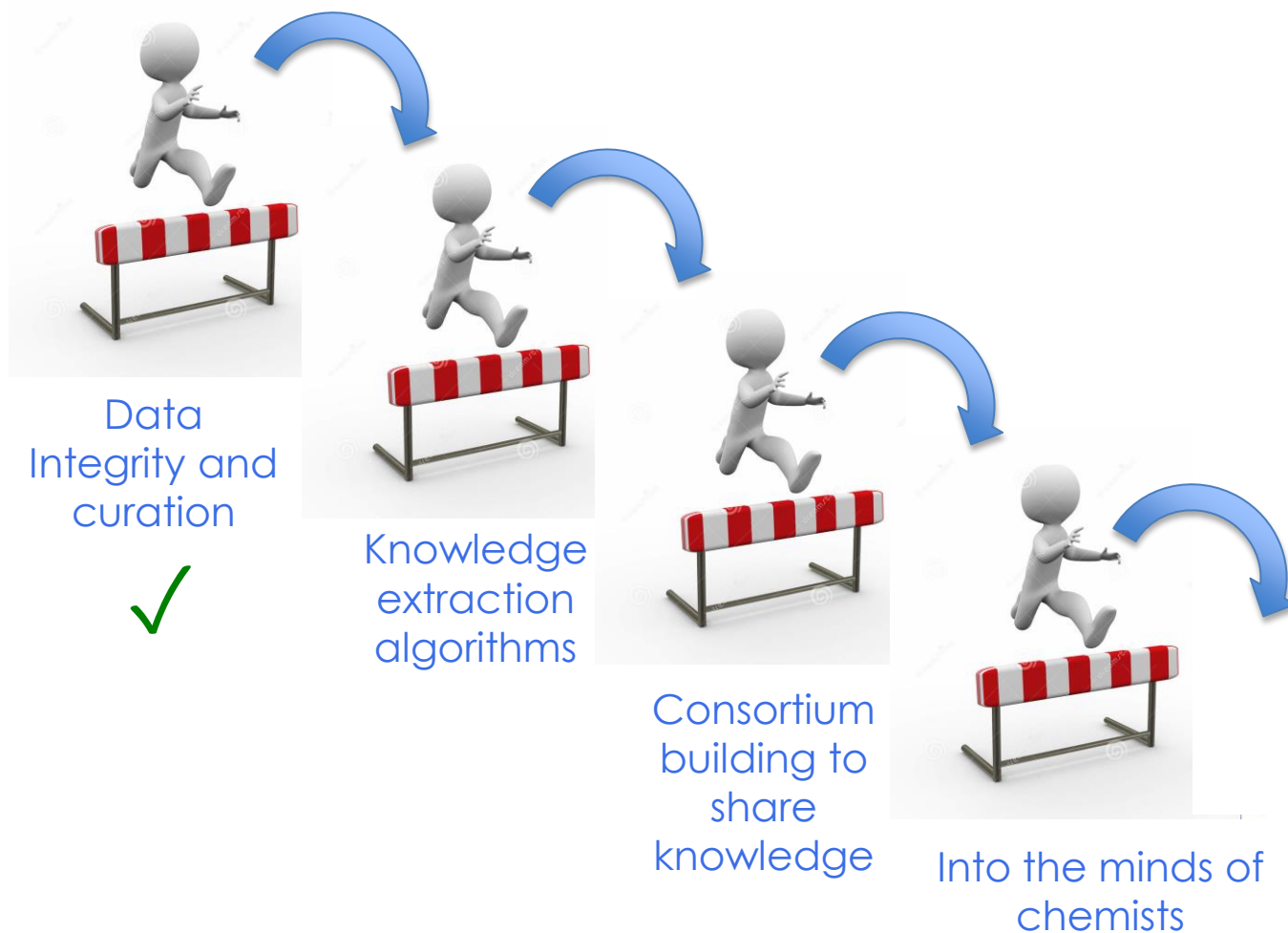
Table. The ChEMBL taxonomy identification is held within the MCPairs database.

idspecies	species	chembl
1	unknown	NULL
2	Rattus norvegicus	10116
3	Canis lupus familiaris	9615
4	Homo sapiens	9606
5	Mus musculus	10090
6	Macaca fascicularis	9541
7	Salmonella enterica subsp. enterica serovar Typhimurium	90371

1962 Species

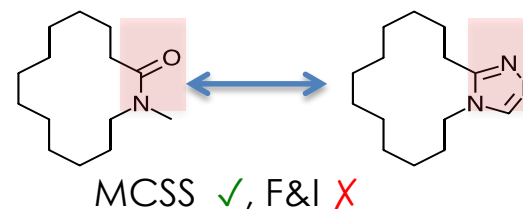
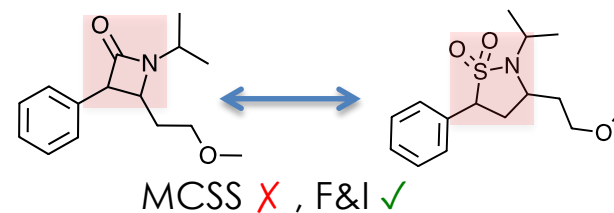
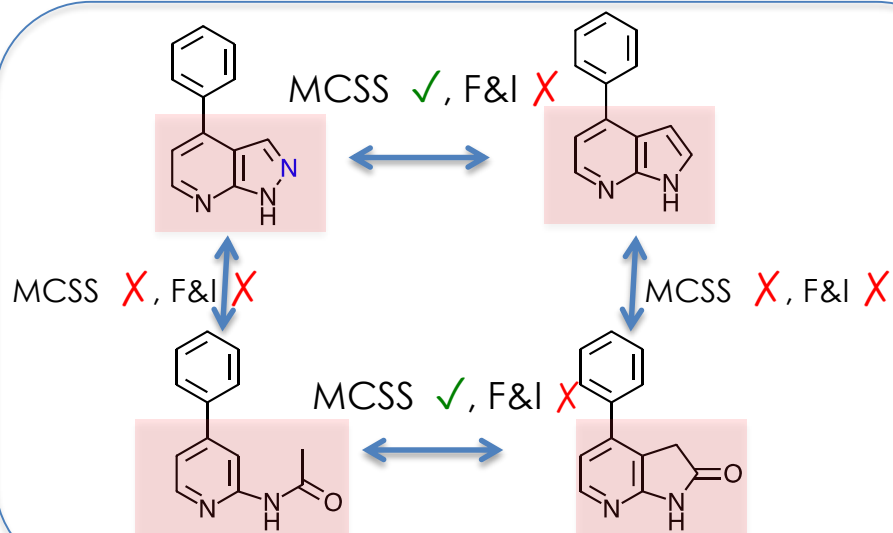
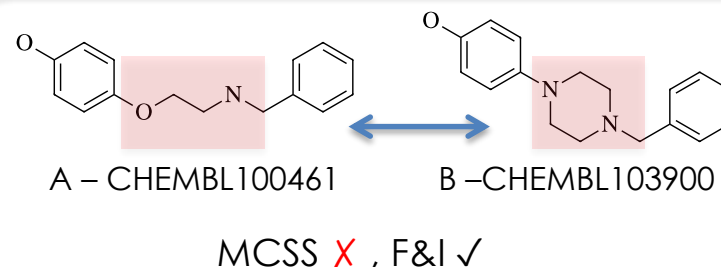
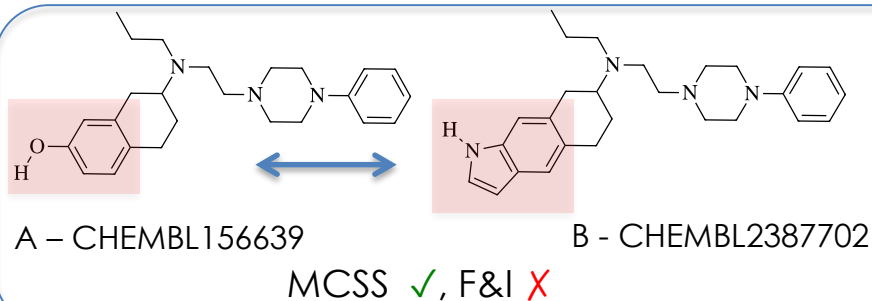
All agreed by Consortium (all in the MCPairs system and documentation)
– use public ontology and taxonomy where possible

Barriers Broken to Sharing Knowledge



Matched pair methodology

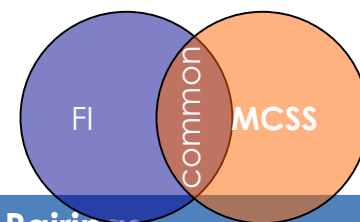
There are two technique – Frag and Index (H/R) and MCSS



The two techniques find different chemistry....

Does the Matched Pair method *really* matter?

Using only one technique will miss between 12% and 56% of pairings

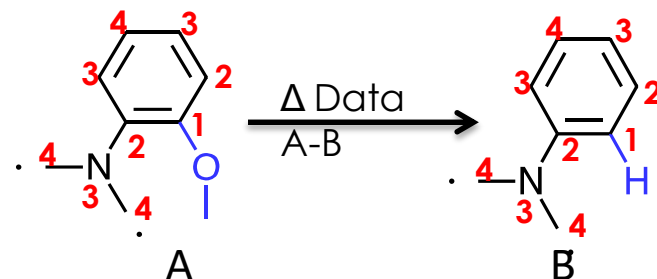
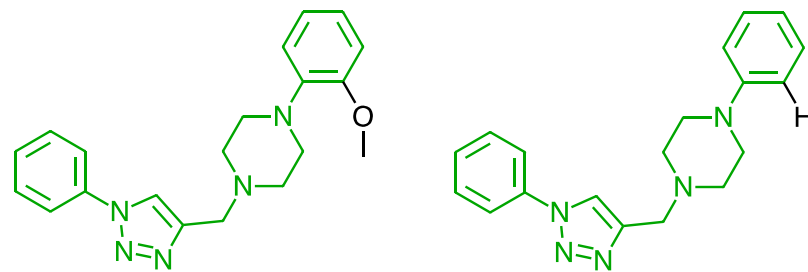


		Pairings				Pairings		
	number of compounds	common	FI only	MCSS only	total	FI only %	common %	MCSS only %
VEGF	4466	14631	17172	14823	46626	37	31	32
Dopamine Transporter	1470	4480	8930	3497	16907	53	26	21
GABAA	848	2500	1722	4205	8427	20	30	50
D2 human	3873	12995	13811	13098	39904	35	33	33
D2 rat	1807	5408	6595	7346	19349	34	28	38
Acetylcholine esterase	383	536	725	1434	2695	27	20	53
Monoamine oxidase	264	653	1156	246	2055	56	32	12
					min	20	20	12
					max	56	33	53

Lukac, I.; Zarnecka, J.; Griffen, E.J.; Dossetter, A.G.; St-Gallay, S.; Enoch, S.; Madden, J.; Leach, A.G. "Turbocharging matched molecular pair analysis; optimizing the identification and analysis of pairs." *J. Chem. Inf. Model.* Submitted

Advanced MMPA with MCPairs

- **Matched Molecular Pairs** – Molecules that differ only by a particular, well-defined structural transformation
- **Transformation with environment capture** – MMPs can be recorded as transformations from A \rightarrow B
- Environment is essential to understand chemistry



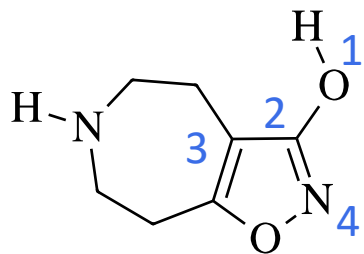
Environment is key and we need to capture it in our chemical encoding...

Griffen, E. et al. Matched Molecular Pairs as a Medicinal Chemistry Tool. *Journal of Medicinal Chemistry*. 2011, **54**(22), pp.7739-7750.

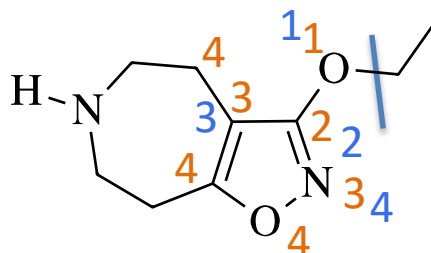
How do we encode the chemical transformation?

- Requirements
 - Lightweight – using as few bytes as possible
 - hashable – allows database indexing
 - Can be used with Chem Toolkits to generate product molecules from chemist's input
 - SMIRKS / Reaction SMARTS (RDKit) fit the bill
 - **Issue** – need to automatically generate SMIRKS from the matched pairs
 - New algorithm required
 - Canonicalisation is required so SMIRKS are consistent from one organisation to another

Standardising Canonicalised SMIRKS



CHEMBL309689



CHEMBL2331793

Orange – atom env radius
Blue – atom map index

Highly specific explicit H

3-Atom rule

[O:1]([H])[c:2]([c:3])[n:4]>>[c:3][c:2]([n:4])[O:1][C]([H])([H])[C]([H])([H])([H])

Key mapped atom

With 4 atom transform environment is more complex

4-Atom rule

[O:1]([H])[c:2]1[c:3]([c:4][o:5][n:6]1)[C:7]([H])([H])>>
[C]([H])([H])([H])[C]([H])([H])[O:1][c:2]1[c:3]([c:4][o:5][n:6]1)[C:7]([H])([H])

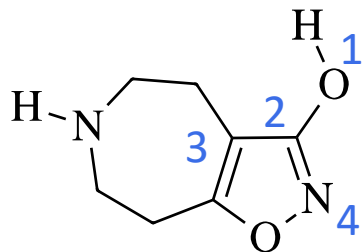
Note Mapped atoms run left to right 1,2,3,4...n

Rule change depending on rule size, environment and symmetry

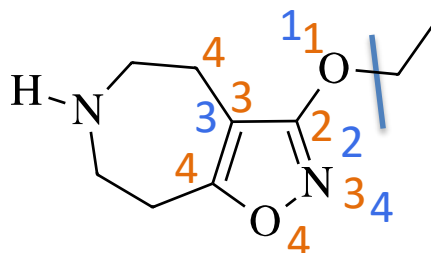
Without explicit H critical information is lost and incorrect products generated

With OpenEye ChemTK SMIRKS operate at 99.2% reliability and functionality

Reaction SMARTs variation (for RDkit)



CHEMBL309689



CHEMBL2331793

Orange – atom env radius
Blue – atom map index

Hydrogens are within SMARTS (note H0 in product)

3-Atom rule $[O;H1:1][c:2]([c:3])[n:4]>>[c:3][c:2]([n:4])[O;H0:1][C;H2][C;H3]$

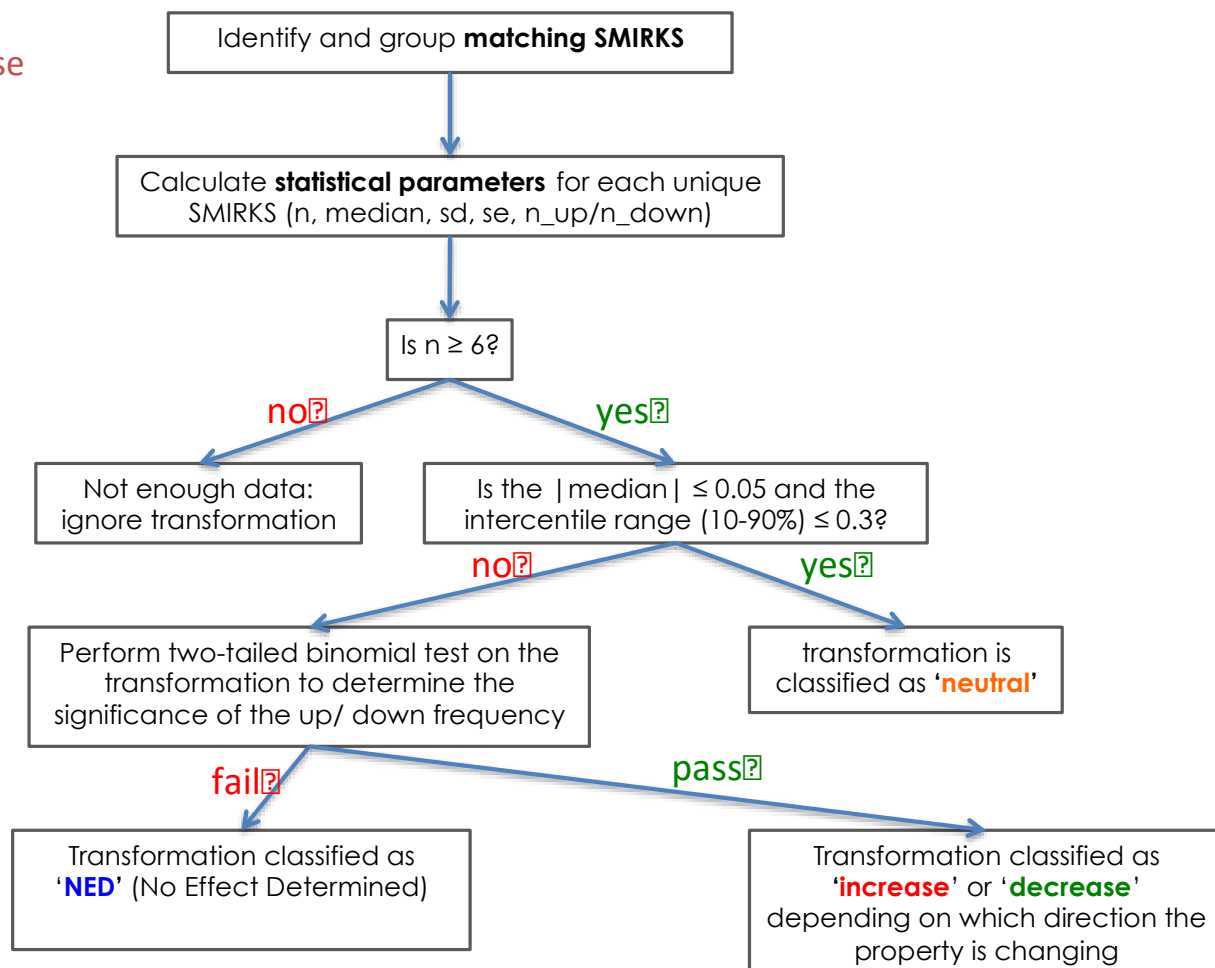
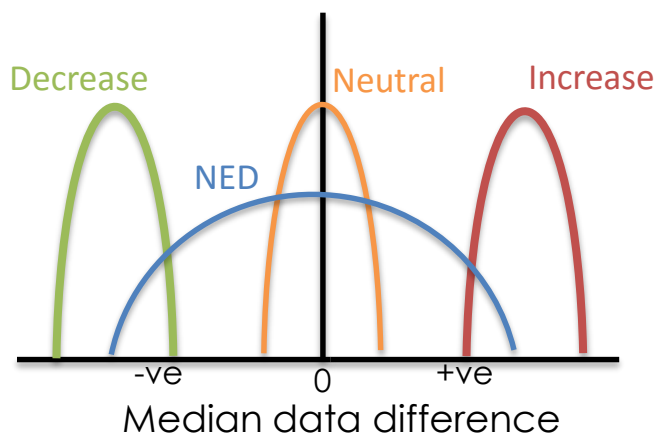
4-Atom rule $[O;H1:1][c:2]1[c:3]([c:4][o:5][n:6]1)[C;H2:7]>>[C;H3][C;H2][O;H0:1][c:2]1[c:3]([c:4][o:5][n:6]1)[C;H2:7]$

RDkit canonical reaction SMARTS work at ~95% of examples.

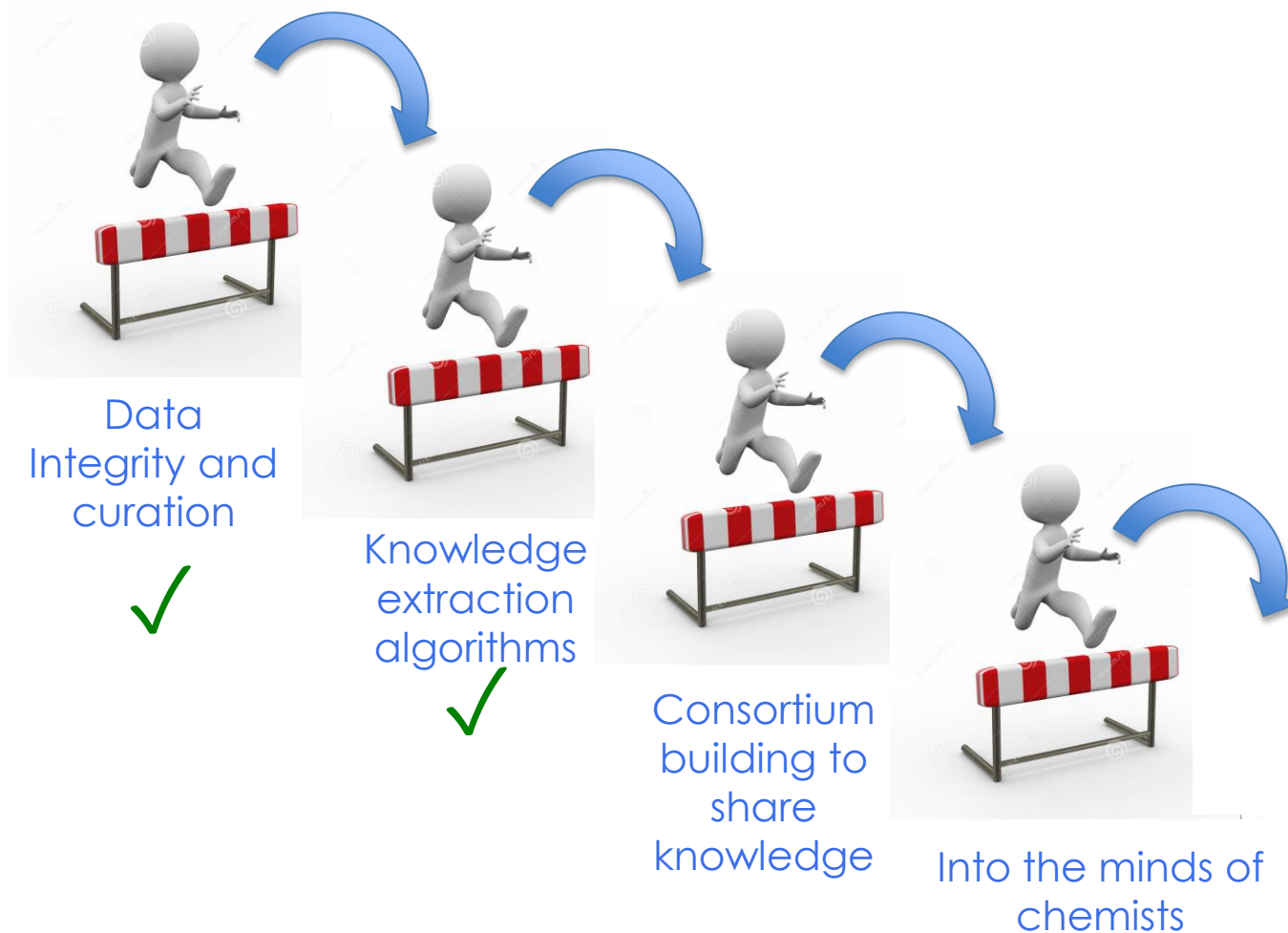
Without the Hydrogen SMARTS in mappings critical chemical information is lost and products are not formed

How do you find Knowledge?

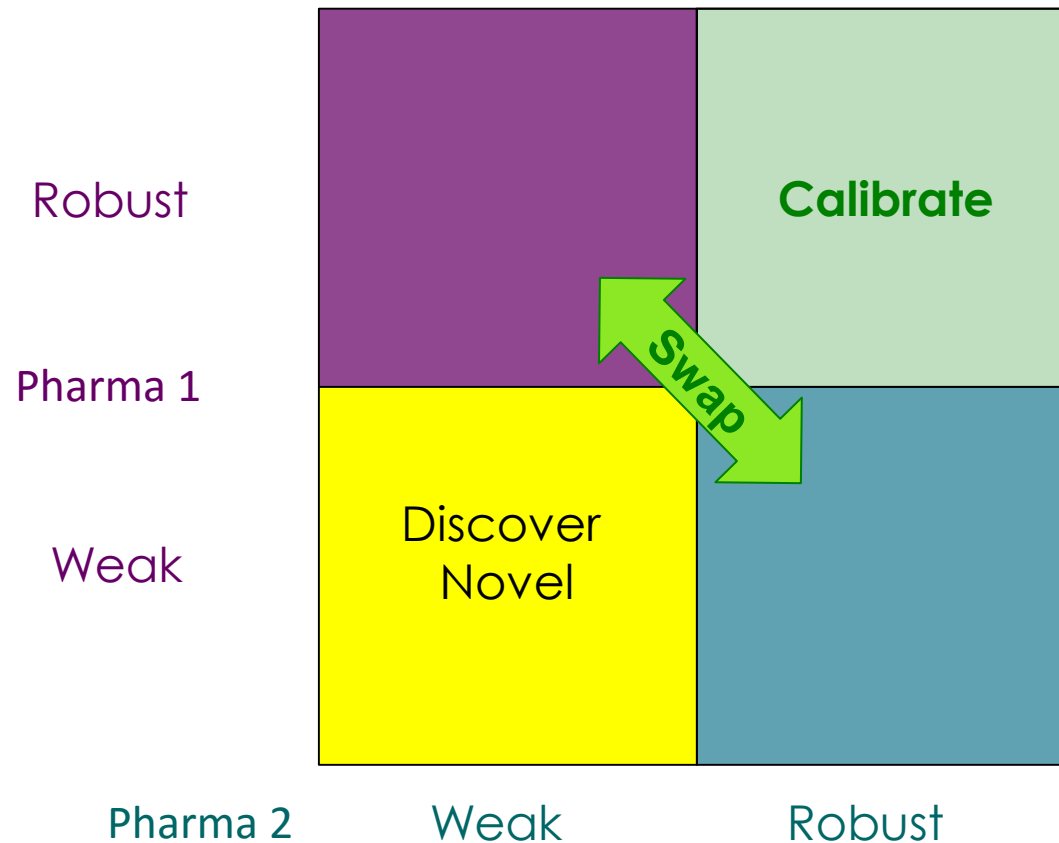
Rule selection 101



Barriers Broken to Sharing Knowledge



Merging knowledge



- Use the transforms that are robust in both companies to calibrate assays.
- Once the assays are calibrated against each other the transform data can be combined to build support in poorly exemplified transforms
- Methodology preceded in other fields

Merging Datasets

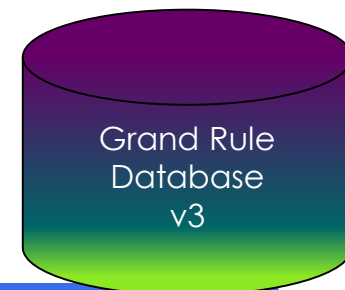
- Datasets are standardized by comparison of transformations shared by contributing companies
- Transformations are examined at the “pair example” level
- Minimum of 6 transformations, each with a minimum of 6 pairs (42 compounds bare minimum) required to standardise
- “calibration factors” extracted to standardize the datasets to a common value – mean of calibration factors 0.94, typical range 0.8-1.2.
- Datasets with too few common transformations have standard compound measurements shared for calibration.



“Blinded” source of transformations

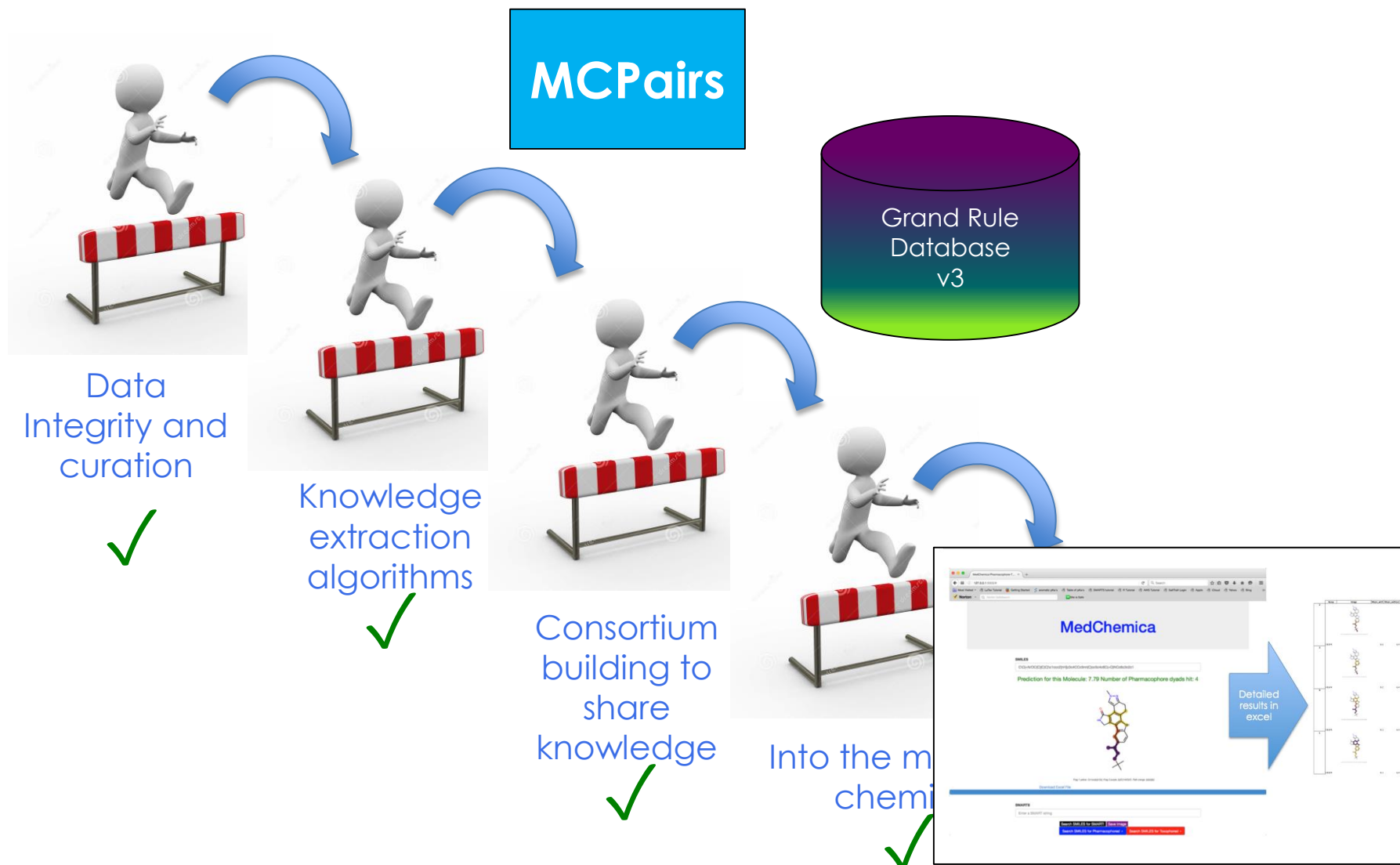
Current Knowledge sets – June 2015

Numbers of statistically valid transforms

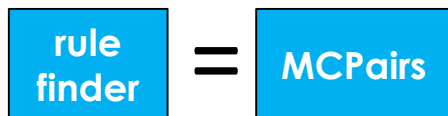
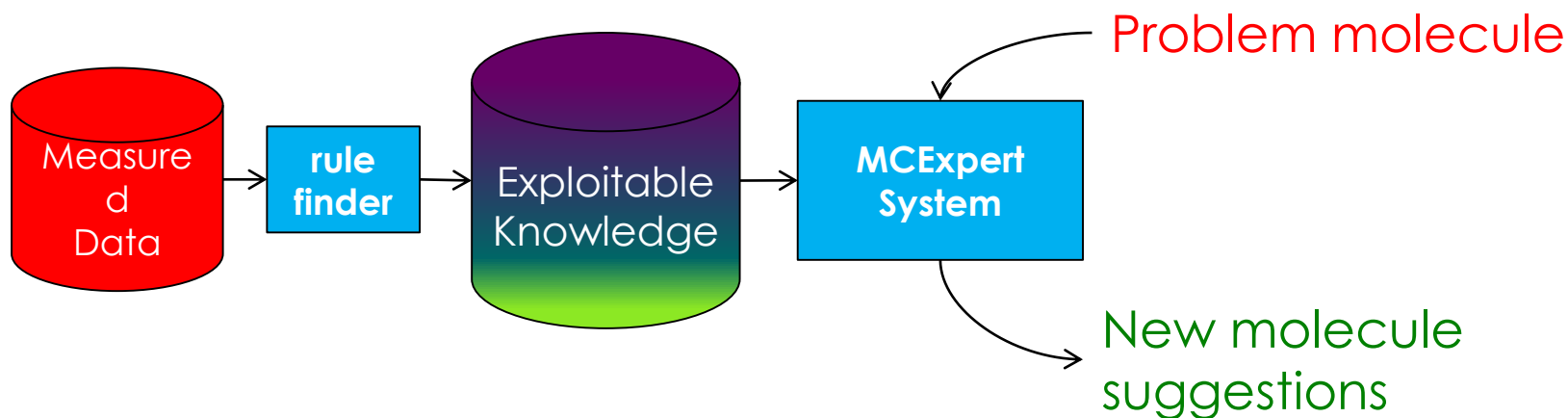


Grouped Datasets	Number of Rules
logD _{7.4}	153449
Merged solubility	46655
In vitro microsomal clearance: Human, rat, mouse, cyno, dog	88423
In vitro hepatocyte clearance : Human, rat, mouse, cyno, dog	26627
MCDK permeability A-B / B – A efflux	1852
Cytochrome P450 inhibition: 2C9, 2D6 , 3A4 , 2C19 , 1A2	40605
Cardiac ion channels NaV 1.5, hERG ion channel inhibition	15636
Glutathione Stability	116
plasma protein or albumin binding Human, rat, mouse, cyno, dog	64622

Barriers Broken to Sharing Knowledge



Exploiting Knowledge for Compound Optimization

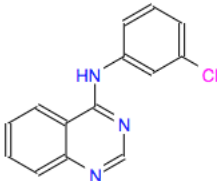


**“..it’s like asking 150 of your peers for ideas in just a few seconds” –
AZ Principal Scientist**

“Its like asking 150 of peers for ideas...”

MedChemica

MCExpert-Lite v0.2-beta



Select Goal ▾

hERG_inhib_pIC50_hum decrease 50 100

Advanced Filters

Molecular Charge Filter ▾

SMILES: Clc3cccc(Nc1ncnc2ccccc12)c3

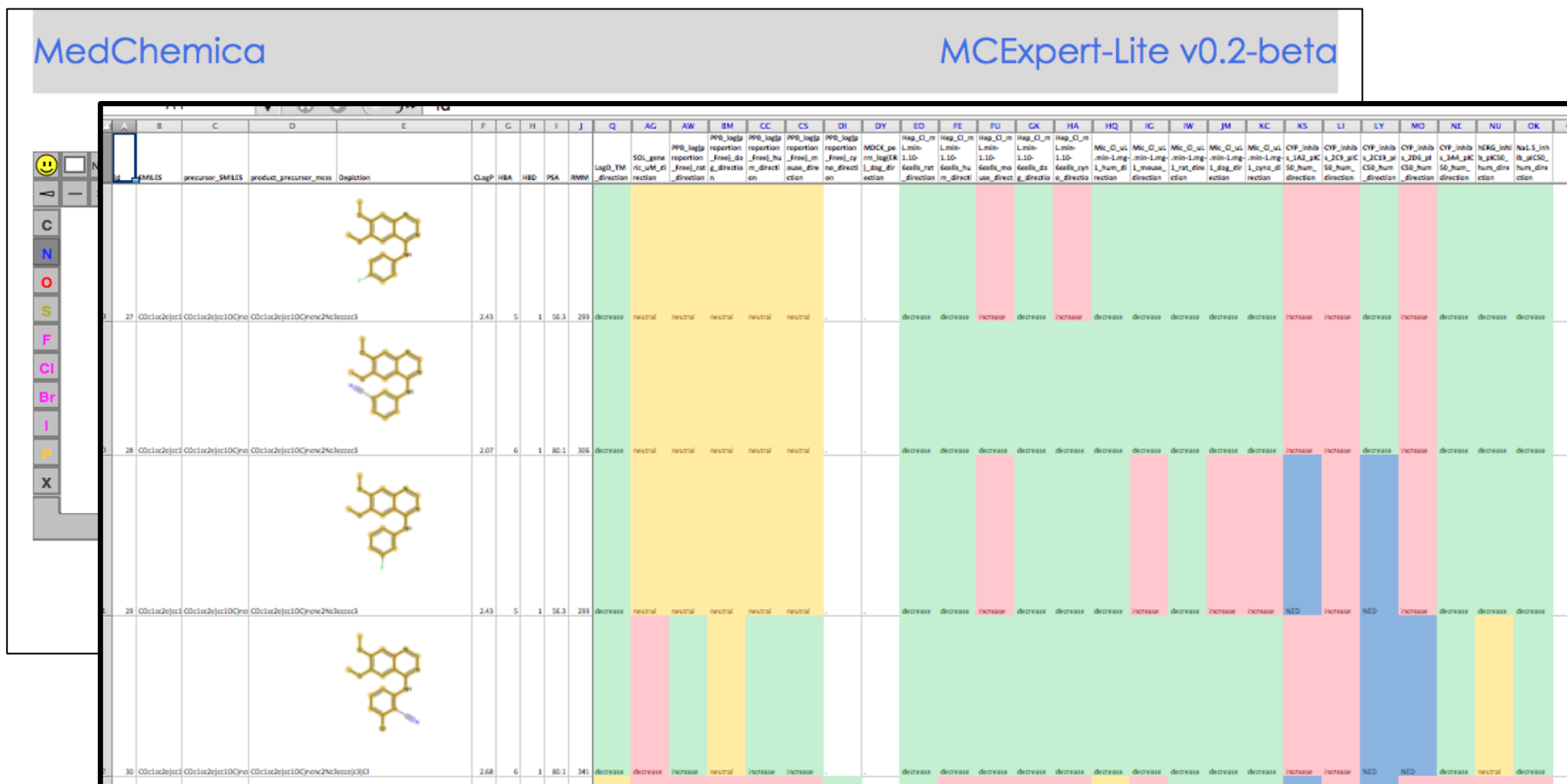
Add Sub-Structure Filter

SUBMIT

Suggested Molecules with “heat map” of Rules for 26 *in-vitro* endpoints
 - “...an MPO Gold Mine” – Roche consortium member

Ask for a demo

“Its like asking 150 of peers for ideas...”



Suggested Molecules with “heat map” of Rules for 26 *in-vitro* endpoints
- “...an MPO Gold Mine” – Roche consortium member

Ask for a demo

Glucokinase Activators

ACS Philadelphia 2016

pEC₅₀: 7.0
logD: 2.9
hERG pIC₅₀: 5.1

ΔpEC
 $n=33$

Roche
Example:

Waring et al. Med. Chem.

MedChemica | 2016

MedChemica | 2016

A Less Simple Example

Increase $\log D$ and gain solu

Solving a ^tBu meta



Property

logD

Log(Solub)

R1

tBu

logD = Kinetic IC50 SS

Q | 2016

	Benchmark compound	Predicted to offer most
R2	tBu	
R1		
	99 392	16 64
	35 128	
	39 445	3 21

- Data shown are Cl_{int} for HLM



Article

pubs.acs.org/imc

Journal of
**Medicinal
Chemistry**

An Orally Bioavailable, Indole-3-glyoxylamide Based Series of Tubulin Polymerization Inhibitors Showing Tumor Growth Inhibition in a Mouse Xenograft Model of Head and Neck Cancer

Helen E. Colley,^{*,†,∇} Munitta Muthana,^{*,∇} Sarah J. Danson,[§] Lucinda V. Jackson,^{||} Matthew L. Brett,^{||} Joanne Harrison,^{||} Sean F. Coole,^{||} Daniel P. Mason,^{||} Luke R. Jennings,^{||} Melanie Wong,^{⊥,∇} Vamshi Tulasi,[⊥] Dennis Norman,[⊥] Peter M. Lockey,[⊥] Lynne Williams,[‡] Alexander G. Dossetter,[#] Edward J. Griffen,^{#,∇} and Mark J. Thompson^{*,||,∇}

[†]School of Clinical Dentistry, University of Sheffield, 19 Claremont Crescent, Sheffield S10 2TA, U.K.

[‡]Department of Oncology, The University of Sheffield, Medical School, Beech Hill Road, Sheffield S10 2RX, U.K.

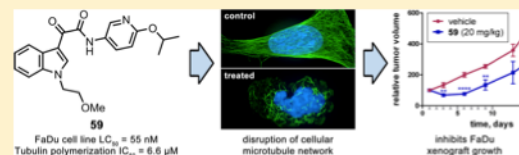
[§]Academic Unit of Clinical Oncology and Sheffield Experimental Medicine Centre, Weston Park Hospital, Whitham Road, Sheffield S10 2SL, U.K.

^{||}Department of Chemistry, University of Sheffield, Brook Hill, Sheffield S3 7HF, U.K.

[†]Charles River, 8–9 Spire Green Centre, Harlow, Harlow, Essex CM19 5TR, U.K.

[†]MedChemica Limited, Ebenezer House, Ryecroft, Newcastle-Under-Lyme, Staffordshire ST5 2BE, U.K.

S Supporting Information



ABSTRACT: A number of indole-3-glyoxylamides have previously been reported as tubulin polymerization inhibitors, although none has yet been successfully developed clinically. We report here a new series of related compounds, modified according to a strategy of reducing aromatic ring count and introducing a greater degree of saturation, which retain potent tubulin polymerization activity but with a distinct SAR from previously documented libraries. A subset of active compounds from the reported series is shown to interact with tubulin at the colchicine binding site, disrupt the cellular microtubule network, and exert a cytotoxic effect against multiple cancer cell lines. Two compounds demonstrated significant tumor growth inhibition in a mouse xenograft model of head and neck cancer, a type of the disease which often proves resistant to chemotherapy, supporting further development of the current series as potential new therapeutics.

Thompson; M.J. *et al J. Med. Chem.*, **2015**, 58 (23), pp 9309–9333

DOI: 10.1021/acs.jmedchem.5b01312

Collaborators and Users



Medicines for Malaria Venture

BLUEBERRY THERAPEUTICS



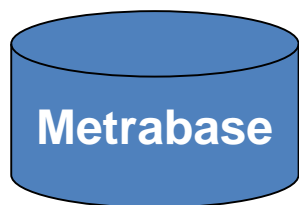
THE UNIVERSITY
of LIVERPOOL



Survey - 17 out of 19 organisations
said the GRD aided project
progression



Can we understand efflux? – MDR1 / PGP



<http://www-metabase.ch.cam.ac.uk>



Pair
Finding

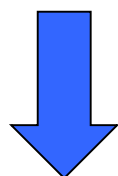


Rule
Finding

1911 compounds: substrate Y/N

Only 826 compound pairs

1 “borderline” rule

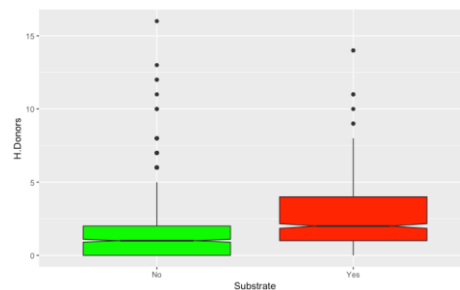
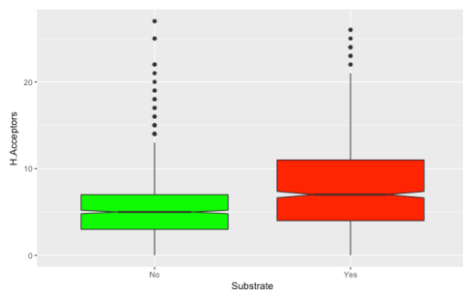
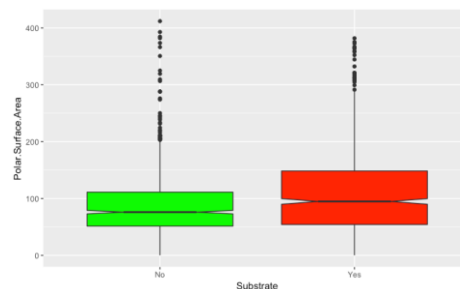
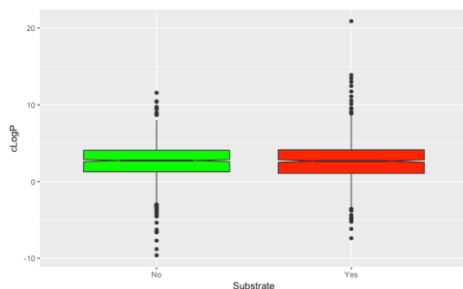
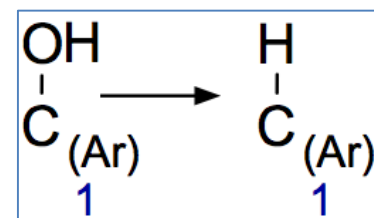


Property analysis

MDR1 substrate = ↑ hydrogen bond donors

↑ hydrogen bond acceptors

↑ PSA



Public transporter data:

- Not quantitative
- Not enough
- Too diverse
- *Trivial conclusions*

See also: Drug Discov Today. 2012 Apr;17(7-8):343-51. doi: 10.1016/j.drudis.2011.11.003

Global Absorption Analysis by MMPA

Combined knowledge from a large number of peer pharma

Secure Analysis of *in-vitro* absorption

Good Absorption improves

- Efficacy
- Safety (lower dose / less off target)

Medicinal Chemistry demands answers

- Low dose oral bio-availability – how?
- MDR1 resistance in oncology
- Brain penetration for CNS diseases
- Only rough and ready “rules” available – trial and error victims

Problem

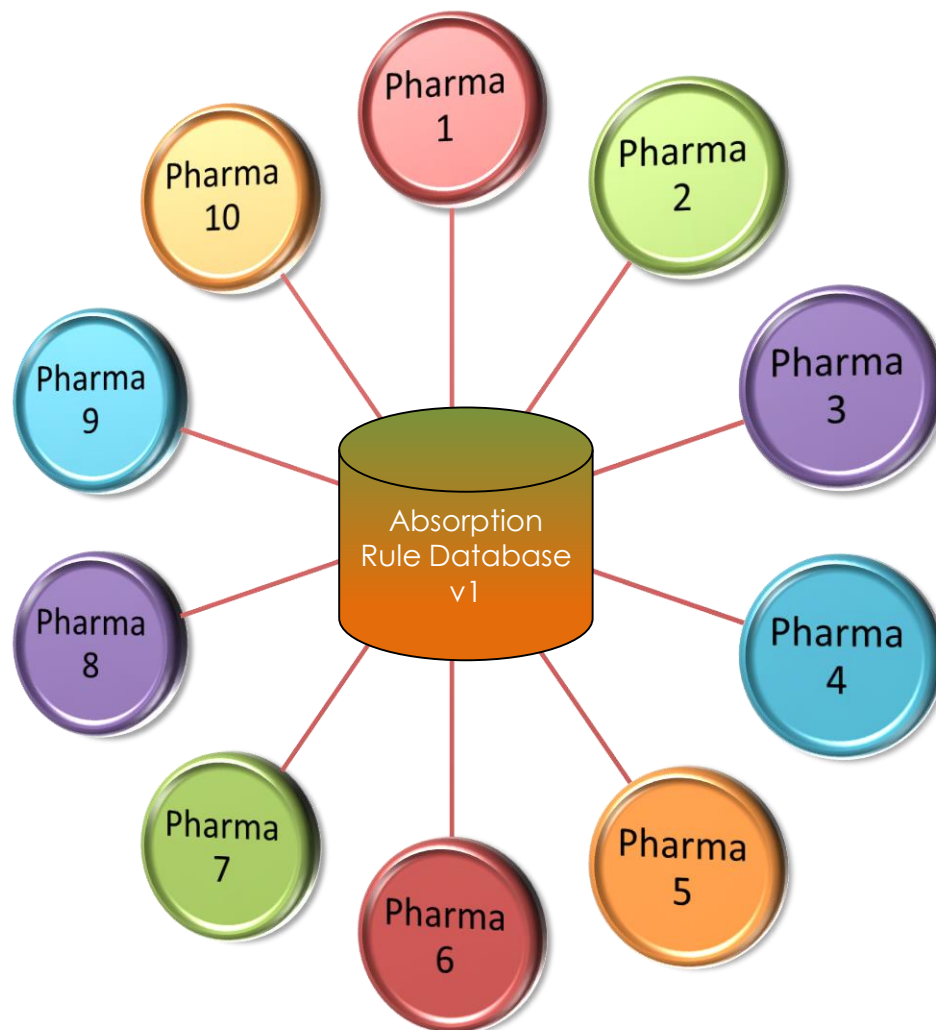
- These expensive assays preclude ANY one company having enough knowledge
- Extreme paucity of data in literature

Solution

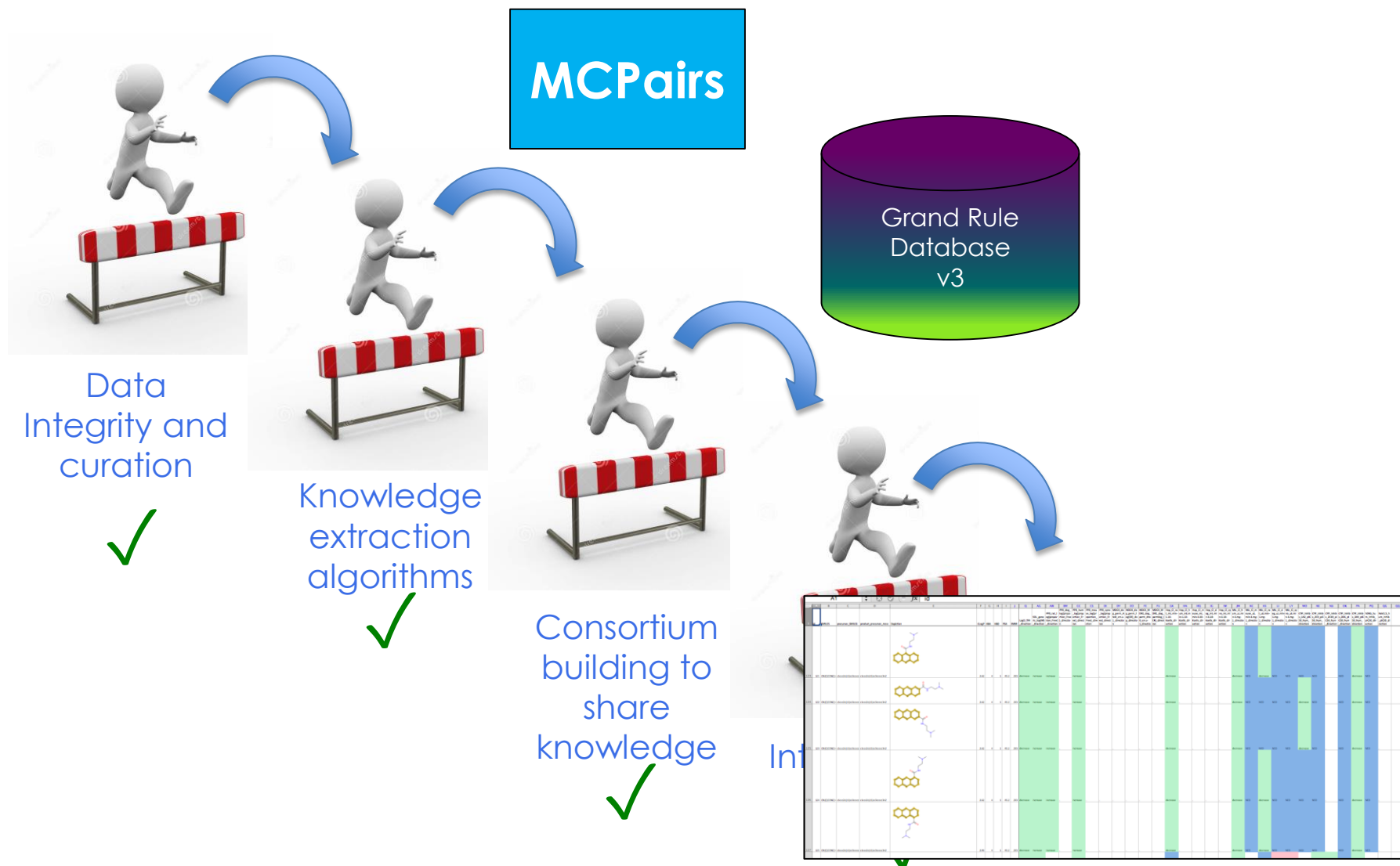
- ≥10 companies worth of data
- A typical pharma has ~10000 results
- MedChemica has the technology, standardization to perform this analysis

MedChemica's offer

\$13600 per organisation to produce a new absorption database



Barriers Broken to Sharing Knowledge



Key findings:

- Secure sharing of large scale ADMET knowledge between large Pharma is possible
- The collaboration generated great synergy
- Standardisation of Units, Species, Assays, MMPA environment, Canonical SMIRKS enabled sharing
- MMP is a great tool for idea generation
- The rules have been used in drug-discovery projects and yields a clear business case for sharing

A Collaboration of the willing

Craig Bruce	OE	Andy Barker	Consulting
John Cumming	Roche	Pat Barton	AZ
David Cosgrove	C4XD	Andy Davis	AZ
Andy Grant★		Andrew Griffin	Elixir
Martin Harrison	Elixir	Phil Jewsbury	AZ
Huw Jones	Base360	Mike Snowden	AZ
Al Rabow	Consulting	Peter Sjo	AZ
David Riley	AZ	Martin Packer	AZ
Graeme Robb	AZ	Manos Perros	Entasis Therapeutics
Atilla Ting	AZ	Nick Tomkinson	AZ
Howard Tucker	retired	Martin Stahl	Roche
Dan Warner	Myjar	Jerome Hert	Roche
Steve St-Galley	Syngenta	Martin Blapp	Roche
David Wood	JDR	Torsten Schindler	Roche
Lauren Reid	MedChemica	Paula Petrone	Roche
Shane Monague	MedChemica	Christian Kramer	Roche
Jessica Stacey	MedChemica	Jeff Blaney	Genentech
		Hao Zheng	Genentech
		Slaton Lipscomb	Genentech
		Alberto Gobbi	Genentech



The background features a large, abstract graphic consisting of several overlapping circles in shades of blue and white. Within these circles are images related to chemistry and technology: a molecular structure with red and blue spheres, a classical building entrance, and a green molecular structure overlaid on a binary code background.

What can Big Data do for Chemistry?

Wednesday 11 October 2017
SCI, London, UK

Organised by SCI's Fine Chemicals Group



The SCI logo consists of a stylized circular icon to the left of the letters 'SCI' in a bold, sans-serif font. Below 'SCI' is the tagline 'where science meets business' in a smaller, lowercase font.

Appendix



Data Integrity and Curation

Structural

- Extensive standards for inclusion of mixtures, chiral compounds, salt forms
- Tautomer and charge state canonicalisation client side
- Automated validation of structures run client side = “clean” comparable structures submitted to pair finding

Measured Data

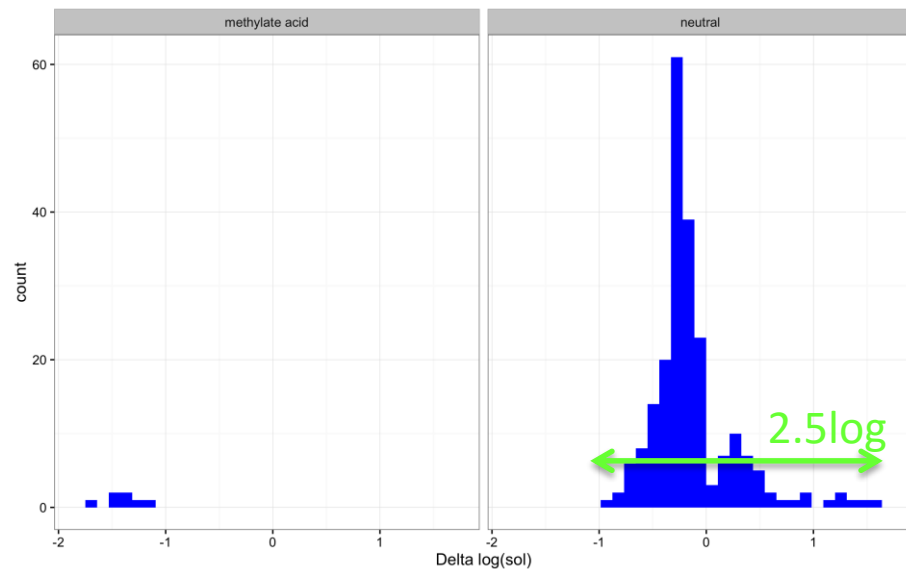
- Assay protocols reviewed prior to merging
- Precise documentation on unit definitions and data reporting standards
- Option to share standard compound measured values
- Automated extensive data validation checks prior to merging data

“client side” = behind Pharma firewall

Environment really matters

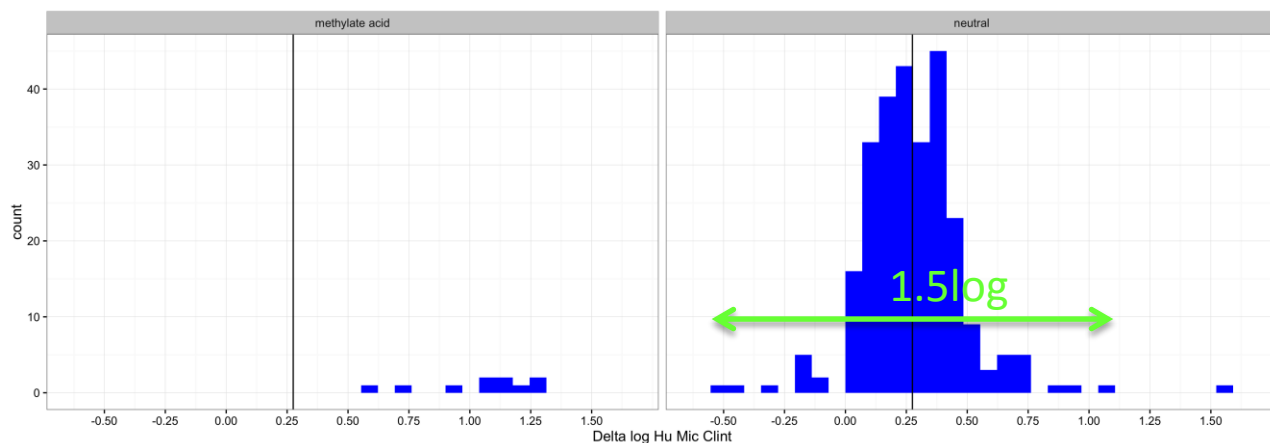
H→Me:

- Median $\Delta\log(\text{Solubility})$
- 225 different environments



H→Me:

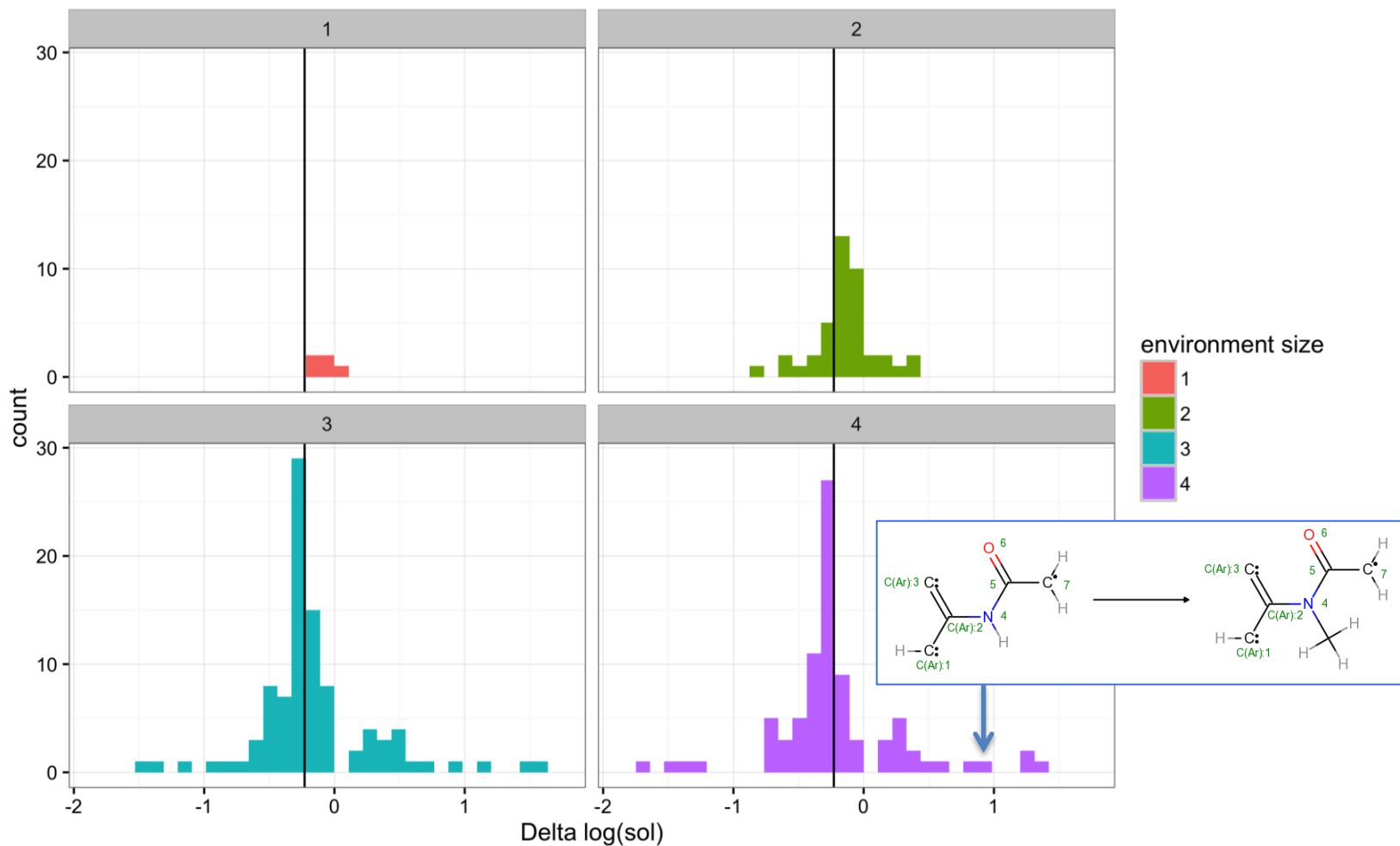
- Median $\Delta\log(\text{Clint})$
Human microsomal clearance
- 278 different environments



More environment = right detail

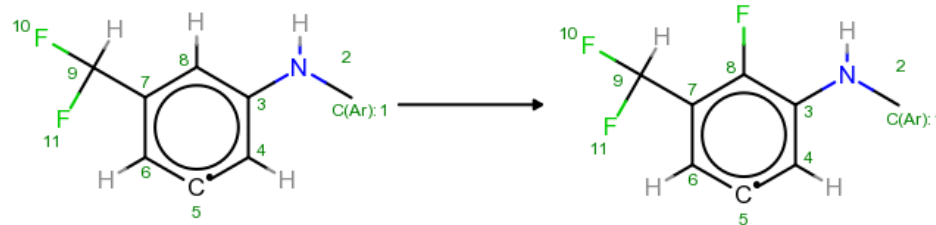
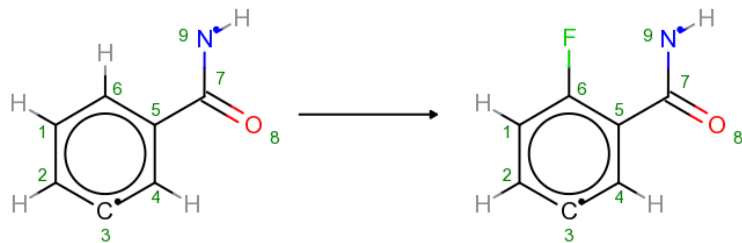
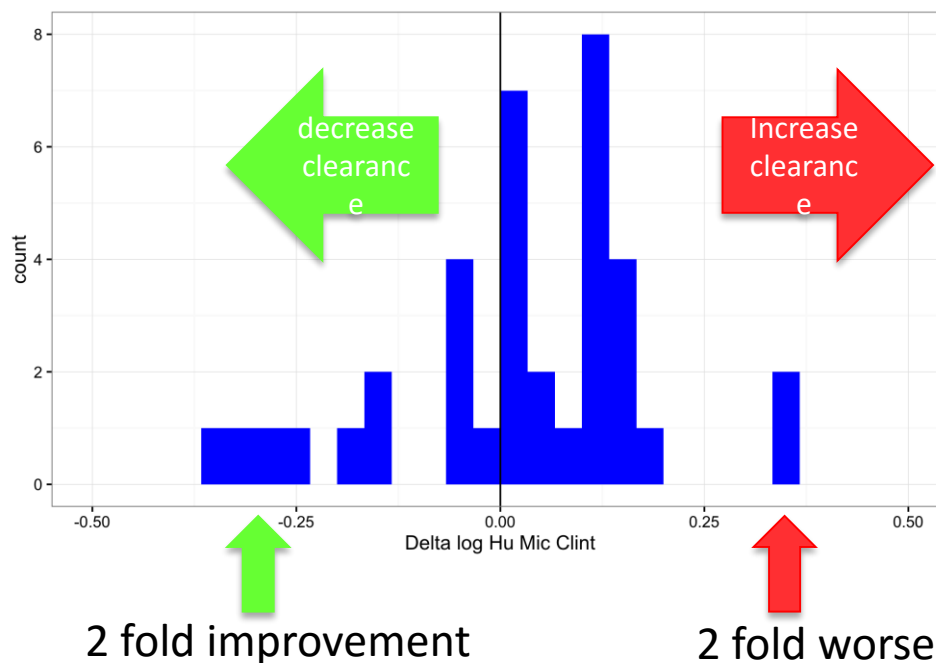
H→Me Solubility:

- 225 different environments

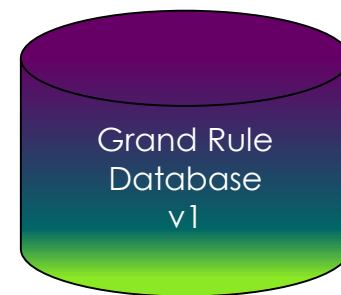


H→F What effect on Clearance?

- Median $\Delta\log(\text{Clint})$ Human microsomal clearance
- 37 different environments



Merging knowledge June 2014



5.8k rules in common (pre-merge) ~ 2%

Pharma 1 100k rules

Pharma 2 92k rules

Pharma 3 37k rules

Merge

New Rules 88k
~26% of total

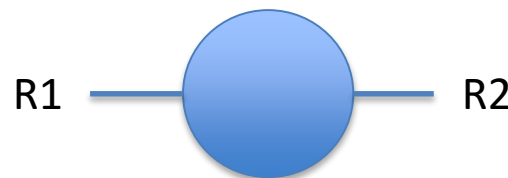
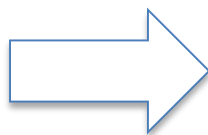
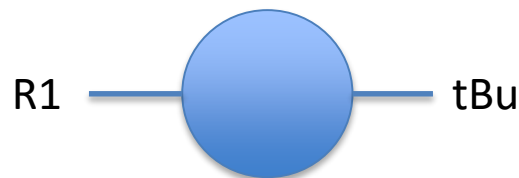
Combining data yields brand new rules

Solving a ^tBu metabolism issue

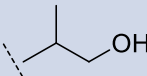
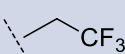
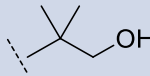
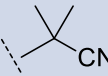
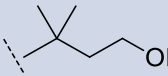

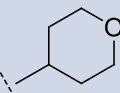
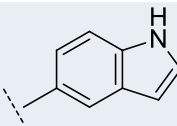
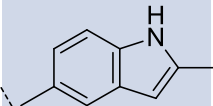
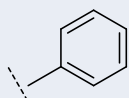


CANCER
RESEARCH
UK

MANCHESTER
INSTITUTE

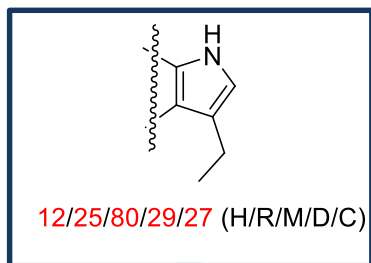


Roger Butlin
Rebecca Newton
Allan Jordan

	Benchmark compound	Predicted to offer most improvement in microsomal stability (in at least 1 species / assay)										
		<div></div>										
R2 R1	tBu				Me				Et	iPr		
	99 392	16 64	78 410	53 550	99 288	78 515	41 35	98 327		92 372	24 247	
	35 128				24 62				60 395			
	39 445	3 21			20 27			57 89		54 89		

- Data shown are Cl_{int} for HLM and MLM (top and bottom, respectively)

The Business Case for Knowledge Sharing



100 cmpds x (\$2K make + \$1K test) = \$ 300 000
8 cmpds x (\$2K make + \$1K test) = \$ 24 000

Enumeration

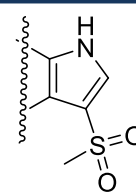
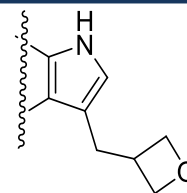
193 compounds
Enumerated

Calculated Property
Docking

8 compounds
synthesized

Objective:
improve
metabolic
stability

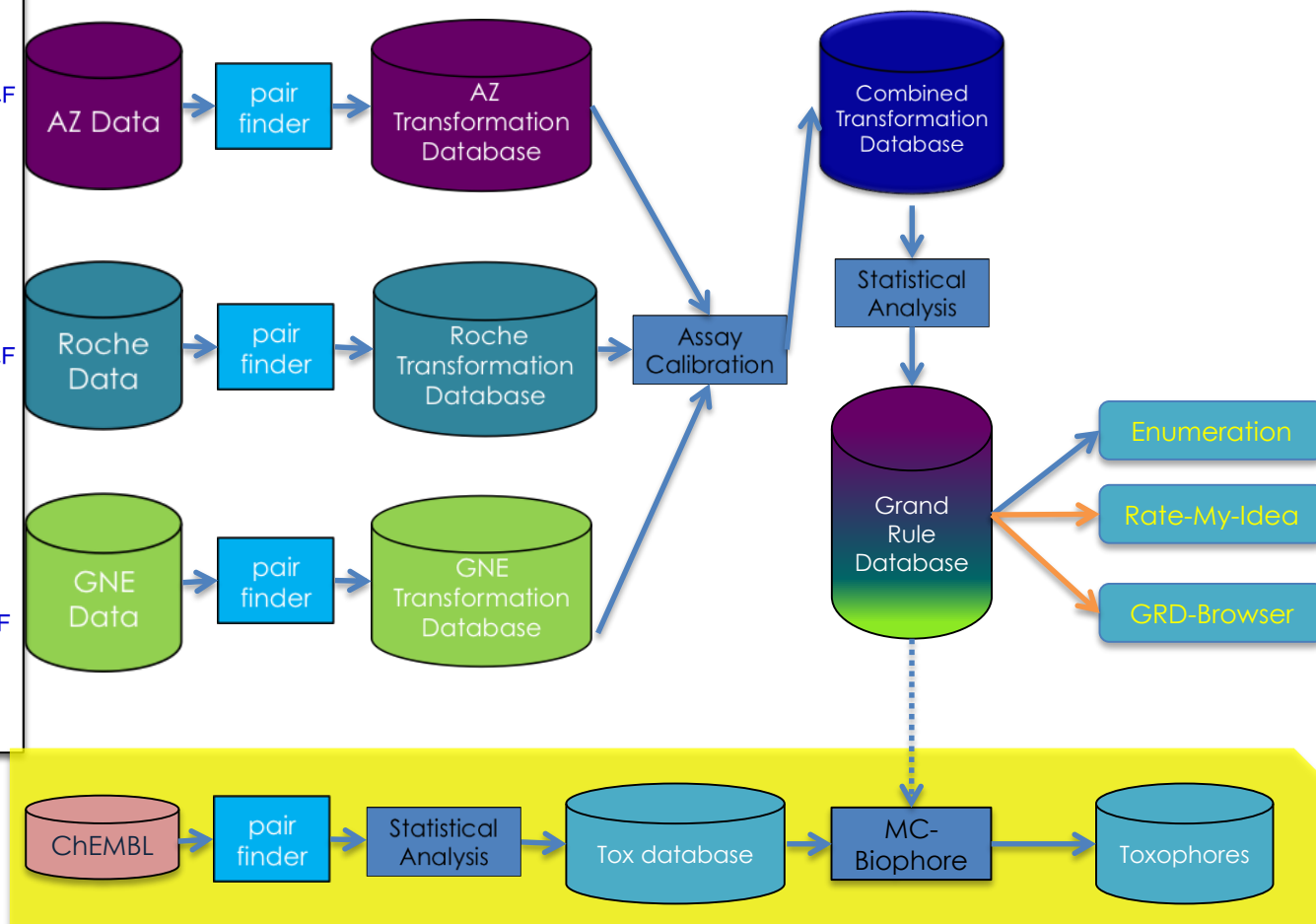
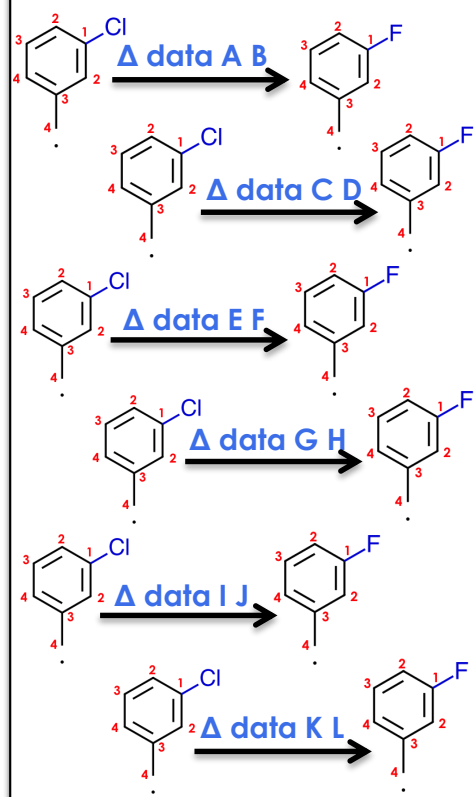
It is not just money, it is actually time
100 cmpds make & test ~ 15 – 25 weeks
8 cmpds make & test ~ 2 – 4 weeks



Matched Molecular Pair Analysis (MMPA) enables SAR sharing

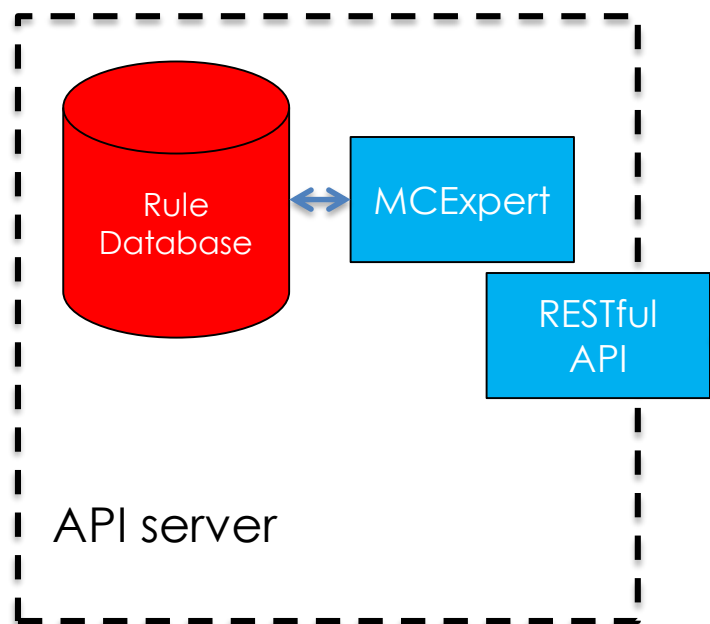
Without sharing underlying structures and data

Chemical Transformations



Integrating knowledge exploitation

..many possible connections
..high Stability and flexibility**



** api.medchemica.com has delivered 18 months of uptime and drives SaltTraX (Elixir)
Approx 50 chemists use the tool



Web tool



visualise



visualise



workflow

Chemistry Shape
and electrostatics

Enhanced
decisions

Chirality MCPairs --chiralON

LHS		RHS	c1ccc(cc1)C2CCCC2C	c1ccc(cc1)C2CCCC	c1ccc(cc1)[C@H]2CCCC	c1ccc(cc1)[C@H]2CCCC	c1ccc(cc1)[C@@H]2CCCC	c1ccc(cc1)[C@@H]2CCCC	c1ccc(cc1)[C@@H]2CCCC	c1ccc(cc1)[C@@H]2CCCC	c1ccc(cc1)[C@@H]2CCCC
			racemic_Cl	racemic_F	RR_cis_Cl	RR_cis_F	SS_cis_absolute	SR_trans_Cl	SR_trans_F	RS_trans_Cl	RS_trans_Cl
c1ccc(cc1)C2CC		racemic_Cl		4							
c1ccc(cc1)C2CC		racemic_F	4								
c1ccc(cc1)[C@H]2CCCC[C@H]2C		RR_cis_Cl				8					
c1ccc(cc1)[C@H]2CCCC[C@H]2F		RR_cis_F			8						
c1ccc(cc1)[C@@H]2CCCC[C@@H]2F		SS_cis_absolute_epimerisation_F									
c1ccc(cc1)[C@@H]2CCCC[C@@H]2Cl		SR_trans_Cl						8			
c1ccc(cc1)[C@@H]2CCCC[C@@H]2F		SR_trans_F					8				
c1ccc(cc1)[C@H]2CCCC[C@@H]2Cl		RS_trans_Cl									

Take home messages

- Systematic “Big Data” analysis of ADMET data yields a system that SUGGESTS solution molecules for medicinal chemists
 - Better decision making / Faster project progression
- Knowledge Sharing IS possible between pharma companies
 - This yields HIGH VALUE databases and VAST new knowledge
- Pharmacophores and toxophores can be extracted from MMPA