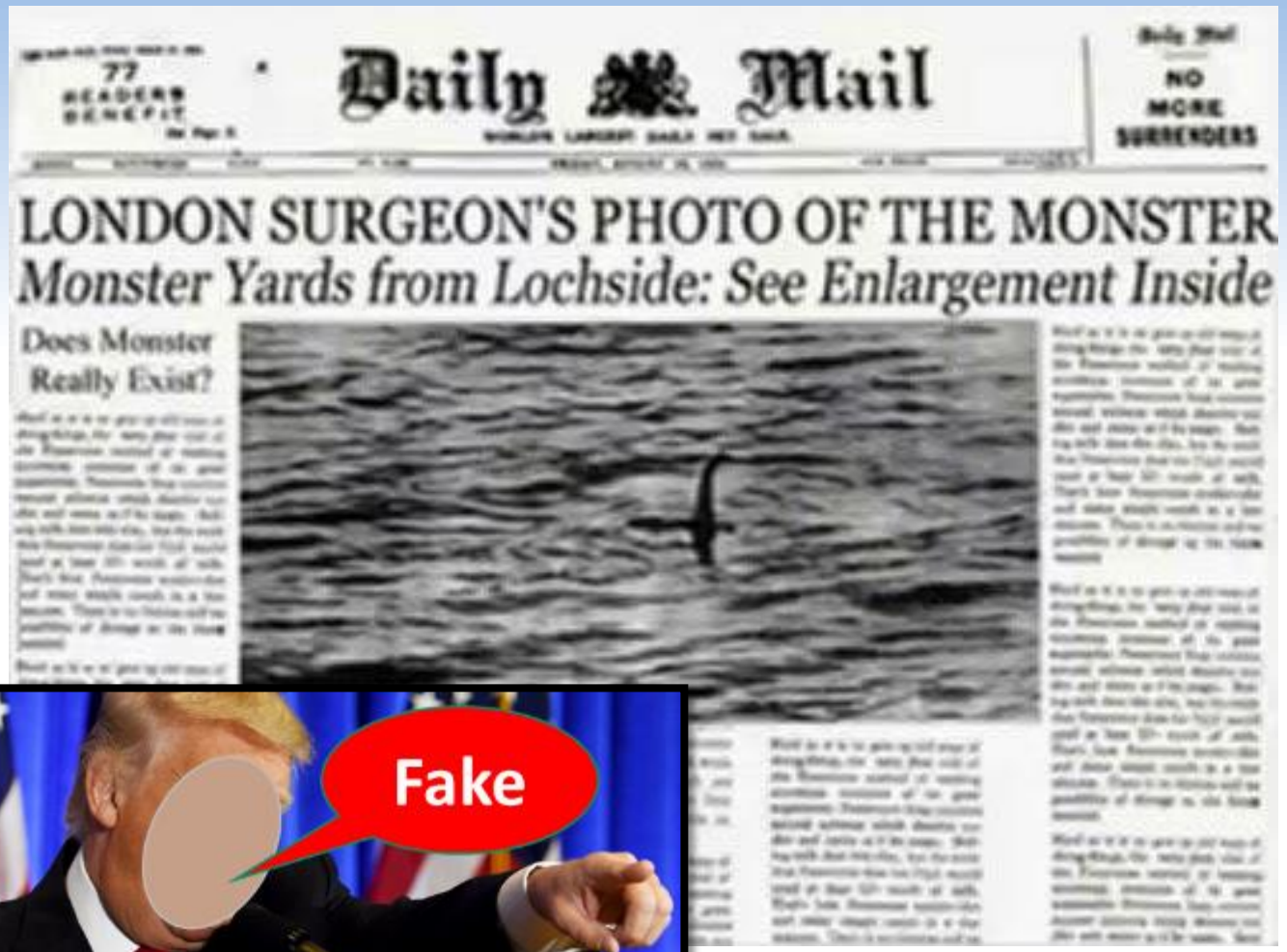
“Data Curation”

The Forgotten Practice in the Era of AI

Pankaj Daga
on behalf of
Cheminformatics Team
Simulations Plus, Inc. Lancaster CA, USA

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Validation is Always Necessary…Be it (Fake) News or Activity Data


Daily Mail

NO MORE SURRENDERS

LONDON SURGEON'S PHOTO OF THE MONSTER

Monster Yards from Lochside: See Enlargement Inside

Does Monster Really Exist?

1934

Roger Patterson and Bob Gimlin (1967)

Disclaimer

- My purpose is NOT to
  - Evaluate/Criticize any public or commercial database
  - Hunt for errors in existing research articles
- My intention is to
  - Suggest how we can utilize existing databases efficiently
  - Outline possible steps to avoid mistakes in literature and/or databases
  - Advocate a strategy for chemical data curation

I solemnly swear that I am up to no good

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Why Data Curation is Necessary?

From: [redacted]
Sent: Tuesday, April 9, 2019 8:46 AM
To: Eric Jamois <eric@simulations-plus.com>
Subject: Re: Access to AP Download Material - ShareFile

Hi Eric

I think it would be useful to have a catch up with you about the prediction accuracy, we have run some test datasets through the ADMEpredictor and we get rather poor correlations, please see the attached word document of the correlations we have done to date with external datasets. If you could set up a meeting with the relevant folks I will send out the invite to the relevant people this end.

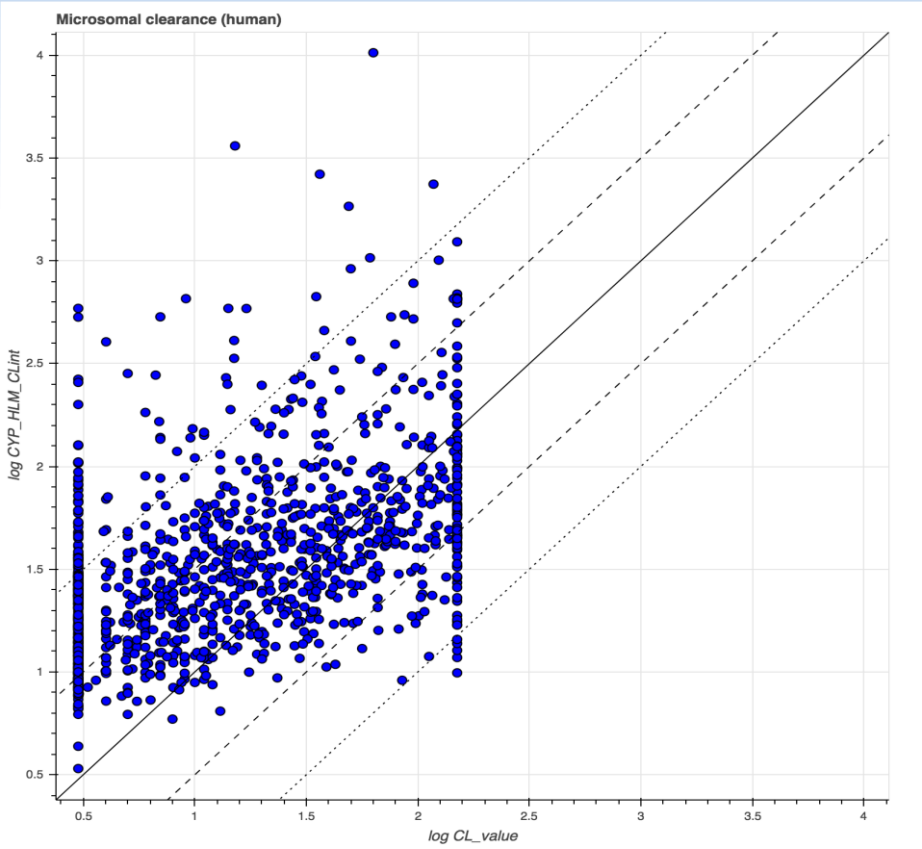
Thanks

[redacted]

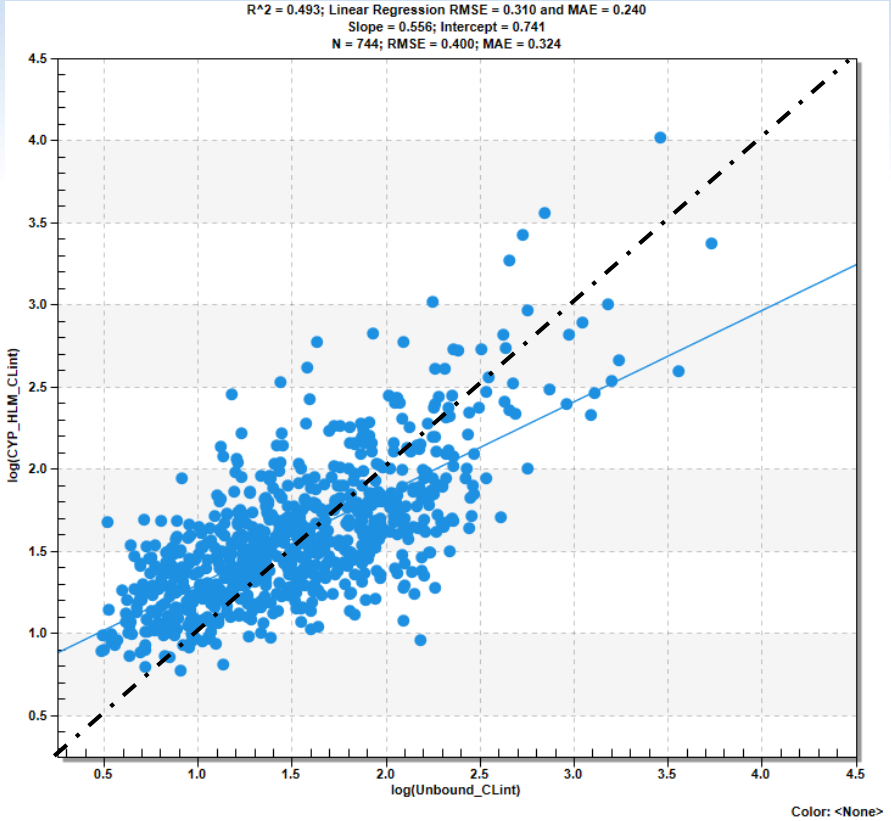
SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Performance of HLM CL_{int} Model

Before Data Curation/Conversions



After Data Curation/Conversions



https://www.ebi.ac.uk/chembl/beta/assay_report_card/CHEMBL3301370/

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

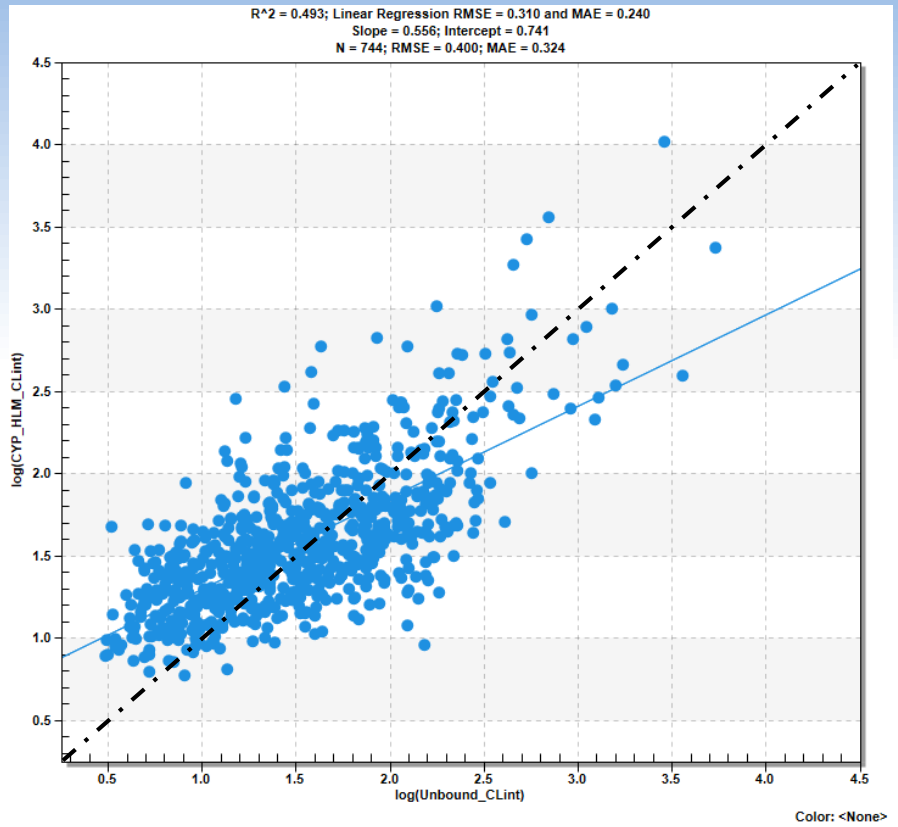
What Did "*The Investigators*" miss???

shown below. Note that *CYP_HLM_CLint*, *CYP3A4_HLM_Km* and *CYP3A4_HLM_CLint* are all corrected for microsomal binding, whereas the other (rCYP) models are not.

- Reported values are bound CLint
- Converted to **Unbound_CLint**

$$Unbound\_CLint = \frac{CL_{int}}{S{+}fumic}$$

- Removed 358 cmpds
  - **84** cmpds have CLint > 150.00
  - **274** cmpds have CLint < 3.00



Had the investigators carried out appropriate data curation & conversions, they would have seen good correlations

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Outline

- What is data validation?
- Where do errors come from?
- How to find them?
- Why should we care about them?

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Outline

- What is data validation?

- Where do errors come from?

- How do we find them?

- Why should we care about them?

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Making Data Fit for Purpose

val·i·da·tion

/ˌvaləˈdāSH(ə)n/

noun

the action of checking or proving the validity or accuracy of something.
"the technique requires validation in controlled trials"

- In computer science, data validation is the process of ensuring data have undergone data cleansing to ensure they have data quality, that is, that they are both correct and useful.

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Validation in (early) Drug Discovery: An Absolute Necessity

- On average, there are two errors per each medicinal chemistry publication
- The overall error rate for compounds can be as high as 8%
- Errors can be introduced during data extraction and digitalization

- For *"accurate and predictive models*, the clean and accurate data is mandatory

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Few Available Bioactivity Databases: Public & Commercial

ChEMBL

LICEPTOR

Reaxys®

GOSTAR
Global Online Structure Activity Relationship Database

Clarivate Analytics

DSSTox

BindingDB

ChemIDplus
A TOXNET DATABASE

PubChem

The usefulness of public data sources is questionable due to lack of the necessary quality control

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Few Available Bioactivity Databases: Public & Commercial

The usefulness of public data sources is questionable due to lack of the necessary quality control

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

If I have seen further it is only by standing on the shoulders of GIANTS

Isaac Newton

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

A Selection of GIANTS

Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research

Denis Fo

Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation

Antony J. Williams[1], Sean Ekins[2] and Valery Tkachenko[1]

Antony J. Williams graduated

correspondence

Curation of chemogenomics data

Denis Fourches[1], Eugene Muratov[2] & Alexander Tropsha[2]

Are the Chemical Structures in Your QSAR Correct?

Douglas Young[a]*, Todd Martin[a], Raghuraman Venkatapathy[b], and Paul Harten[a]

Tales from the war on error: the art and science of curating QSAR data

Marvin Waldman[1] · Robert Fraczkiewicz[1] · Robert D. Clark[1]

Besides chemical structure information, quality of QSAR models also strongly depends on the target biological data.

Analysis of Commercial and Public Bioactivity Databases

Pekka Tiikkainen*,† and Lutz Franke†

Parallel Worlds of Public and Commercial Bioactive Chemistry Data

Miniperspective

Christopher A. Lipinski,† Nadia K. Litterman,‡ Christopher Southan,§ Antony J. Williams,‖ Alex M. Clark,⊥ and Sean Ekins*,‡,#

Estimating Error Rates in Bioactivity Databases

Pekka Tiikkainen,*,† Louisa Bellis,‡ Yvonne Light,‡ and Lutz Franke†

The Experimental Uncertainty of Heterogeneous Public K_i Data

Christian Kramer,*,† Tuomo Kalliokoski,*,† Peter Gedeck, and Anna Vulpetti

Activity, assay and target data curation and quality in the ChEMBL database

George Papadatos[1] · Anna Gaulton[1] · Anne Hersey[1] · John P. Overington[1]

SimulationsPlus

SCIENCE + SOFTWARE = SUCCESS

Outline

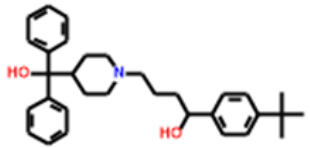
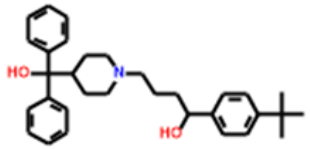
- What is data validation?
- **Where do errors come from?**
- How to find them?
- Why should we care about them?

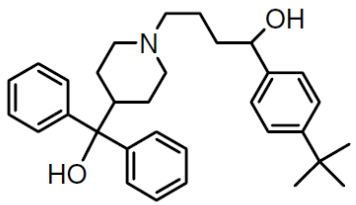
SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Sources of Errors: Data Extraction

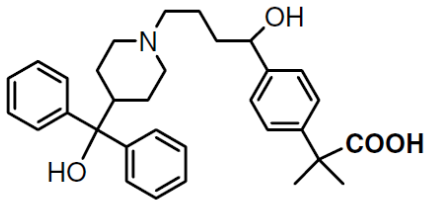
Unit inconsistencies is very common during data digitization

Difference of 3/6 orders of magnitudes suggests the error during digitization

	Structure	Identifier	Standard.Relation	Standard.Units	IC50	Document
9		Cmpd X	=	nM	56000.000	https://www.ebi.ac.uk/chembl/document_report_card/CHEMBL1139687
10		Cmpd X	=	nM	56.000	https://www.ebi.ac.uk/chembl/document_report_card/CHEMBL1145042



Terfenadine **41**
hERG IC50 = 56 nM



Fexofenadine **42**
hERG IC50 = 23uM

Zhu et al. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 5507

Table 1. Comparison of the HERG Channel Affinity to That of the Intended Pharmacological Target for Several Drugs

drug	target affinity	HERG IC_{50}	comment
terfenadine	58 nM (histamine H1 K_i)	56 nM	withdrawn
astemizole	3 nM (histamine H1 K_i)	0.9 nM	withdrawn
cisapride	29 nM (serotonin 5HT4 K_i)	47 nM	withdrawn
sertindole	0.6 nM (serotonin 5HT2A K_i)	3 nM	withdrawn
thioridazine	27 nM (dopamine D2 K_i)	191 nM	black box[a]
pimozide	12 nM (dopamine D2 K_i)	18 nM	TDP[b]
grepafloxacin	up to 2.4 μM (bacterial MIC[c])	50 μM	withdrawn

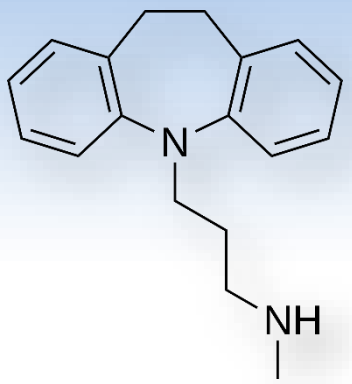
Pearlstein et al. *J. Med. Chem.* **2003**, *46*, 2017

Sources of Errors: Data Extraction

Commercial Databases

Bioavailability of Desipramine: 0.15

Looks quite realistic …..



Ranges for Fraction bioavailable
0.0 – 1.0

SAR Data | ADME Data | Metabolite Data | Toxicity Data | Clinical Data

- Activity Information

S.No	REF ID	Source	Activity Type	Activity	Enzyme Cell Assay	Parameter	Assay Type
1	36683	HUMAN	Fu	= 0.15		Bioavailability	A
2	36683	HUMAN	Vdss	= 20 L/Kg		Distribution	D

Lombardo et al. *J. Med. Chem.* **2004**, *47*, 1242

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

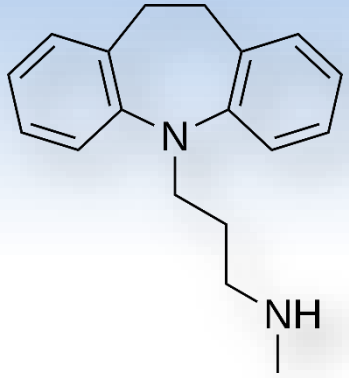
Sources of Errors: Data Extraction

Commercial Databases

Bioavailability of Desipramine: 0.15

Looks quite realistic…..

UNTIL we notice the "**Assay Methods**"



Ranges for Fraction Unbound
0.0 – 1.0

Ranges for Fraction bioavailable
0.0 – 1.0

SAR Data | ADME Data | Metabolite | Toxicity Data | Clinical Data

- Activity Information

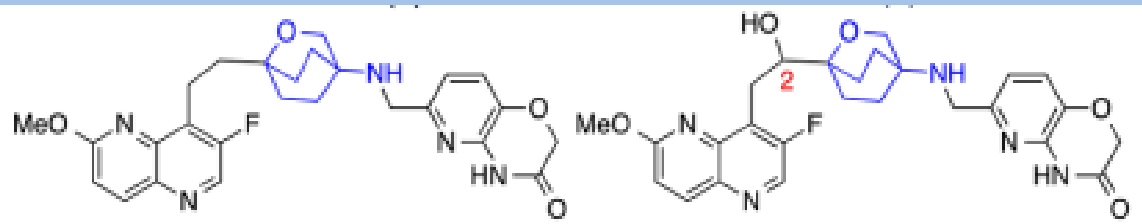
S.No	REF ID	Source	Activity Type	Activity	Enzyme Cell Assay	Parameter	Assay Type
1	36683	HUMAN	Fu	= 0.15	Fraction of compound unbound in human plasma was determined	Bioavailability	A
2	36683	HUMAN	Vdss	= 20 L/Kg	Volume of distribution of the compound at steady state was determined in human	Distribution	D

Lombardo et al. *J. Med. Chem.* **2004**, *47*, 1242

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

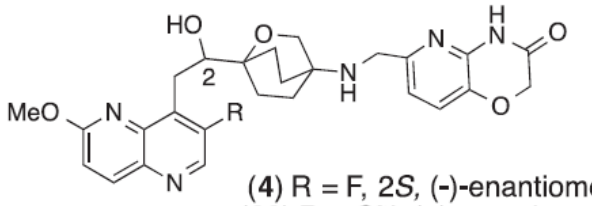
Sources of Errors: Original Research Articles

Singh et al, ACS Med. Chem. Lett. **2014**, *5*, 609



agents. We evaluated hERG activity in a functional automated patch clamp assay (see Supporting Information for Methods). In this assay, AM8085 showed an IC_{50} of 0.6 μM. The 2*S*-hydroxy group of AM8191 improved the polarity and attenuated the hERG activity (IC_{50} = 18 μM). However, more than an order of magnitude attenuation of hERG activity may be required for a clinical development compound.

Singh et al, Bioorg. Med. Chem. Lett. **2015**, *25*, 2409



(**4**) R = F, 2*S*, (-)-enantiomer
(**88**) R = CN, (+)-enantiomer
(**89**) R = CN, (-)-enantiomer

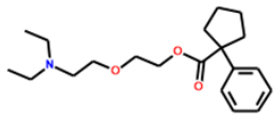
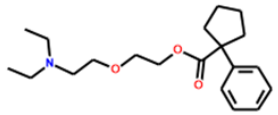
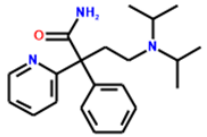
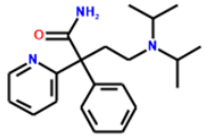
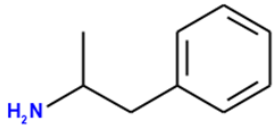
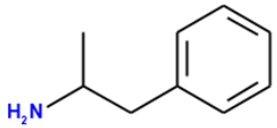
List	R	SaS	SaMR	Sp	Ef	Ec	Ab	Pa	hERG binding (IC_{50}, nM)	PX hERG (IC_{50}, nM)	$c\log D_{7.4}$
4	F	0.02	.06	.05	0.5	2	0.5	8	26.00	18.00	1.9
88	CN	0.031	0.031	0.25	1	4	0.5	8	2.13	2.00	1.3
89	CN	0.063	0.031	0.25	2	4	1	16	2.58	NT	1.3

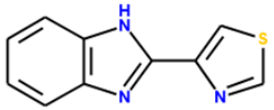
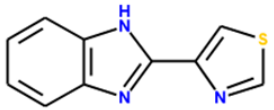
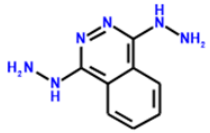
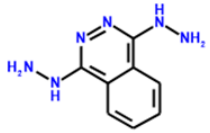
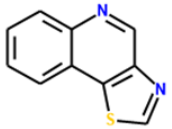
Our model suggested that the correct hERG IC_{50} of AM8191 is **18 nM** and NOT **18µM**

Sources of Errors: Database users

Different salts may (or may not) have different activities/toxicities

One cannot necessarily compare activities/toxicities of different salts

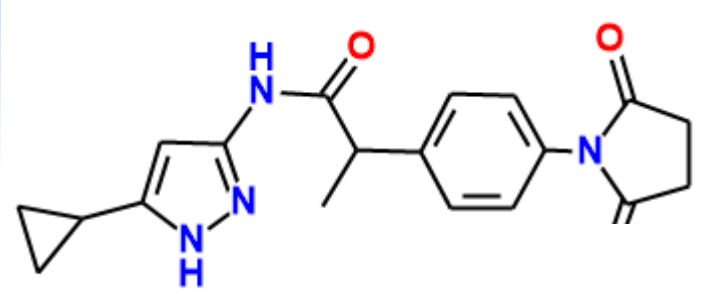
Structure	Identifier	Chemical_Name	Salt_Solvent	LD50_mgkg
	23142-01-0	Carbetapentane citrate [NF]	OC(CC(O)=O)(CC(O)=O)C(O)=O citrate	810 810
	77-23-6	Carbetapentane	? --	150 150
	22059-60-5	Disopyramide Phosphate	OP(=O)(O)O phosphate	880 880
	3737-09-5	Disopyramide	? --	333 333
	139-10-6	Dextroamphetamine Phosphate	OP(=O)(O)O phosphate	302 302
	51-64-9	Dextroamphetamine	? --	38 38

Structure	Identifier	Chemical_Name	Salt_Solvent	LD50_mgkg
	28558-32-9	Thiabendazole hypophosphite	P(O)=O Hypo-phosphite	3100 3100
	148-79-8	Thiabendazole	? --	2080 2080
	7327-87-9	Dihydralazine sulfate	S(O)(O)(=O)=O sulfate	400 400
	63868-75-7	Dihydralazine hydrochloride	Cl --	350 350
	111199-29-2	Thiazolo(4,5-c)quinol... monoethanesulfonate	S(O)(=O)(=O)CC Monoethyl-sulfonate	290 290
	111199-28-1	Thiazolo(4,5-c)quinol... monohydrochloride	Cl --	350 350

Sources of Errors: A(B)CD

Automated (and Blind) Compilation of Data

It is tempting to automate curation itself by accepting as correct (…….) but the potential for (false) positive reinforcement (….) is dangerously high.



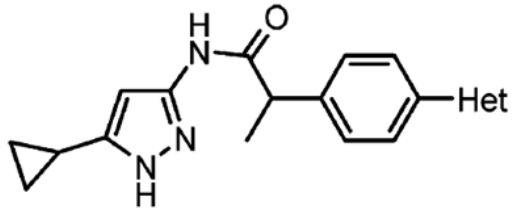
21

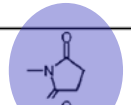
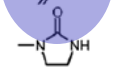
The Compound **21** shows extracellular double bond in original article

Transcribed likewise in the database

Always "jumped-out" as an outlier in our in-house Rat PPB model. Hence needed careful review.

Table 3. SAR of 3-(4-Heterocycl-1-yl)phenyl-acetamido-5-cyclopropyl-1*H*-pyrazoles (**21**–**32**)



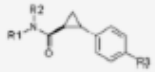
Entry	Het	α-methyl configuration	CDK2/cyclin A (IC_{50}; nM)[a]	A2780 (IC_{50}; nM)[a]	Caco-2 Permeability	Solubility (μM; buffer pH 7)	Plasma Protein Binding (%)
21		R,S	77	>10,000	Moderate	220	48
22		R,S	12	2,250	Moderate	224	74
23		R	455	13,200	Moderate	220	74
24		S	2	1,270	Moderate	>225	74
25		R,S	17	4,540	Low	201	67
16		R,S	150	6,400		222	67

Waldman et al, *J Comput Aided Mol Des*, **2015**, *29*, 897

Pevarello et al, *J. Med. Chem.* **2005,** *48*, 2944

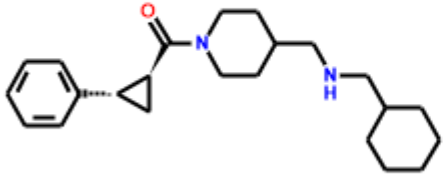
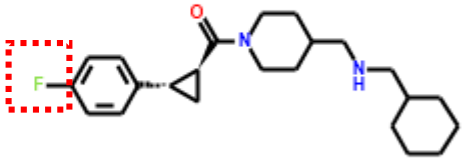
Sources of Errors: A(B)CD

Table 1 SAR and key *in vitro* properties of phenylcyclopropylcarboxamide analogs



Compound	R1–N(R2)	R3	σ1 pK_i[a]	σ2 pK_i[b]	Off-targets profiling	log *D* [ACD_log *D*][f]	LLE [cLLE][d]	pK_a [ACD_pK_a][f]	Cl_{int} (rat)[f]
(±)-1		H	6.7 ± 0.38	*nt*	**Ca^{2+}**, 5HT2a, α1[g]	1.3	5.4	10.1	*nt*
(+)-1		H	6.8 ± 0.38	6.6	**5HT2a**, Na^+, α1, α2c, α2b, Ca^{2+}[h]	1.1	5.7	10.4	147
(−)-1		H	6.8 ± 0.27	*nt*	Na^+[g]	1.1	5.7	10.4	*nt*
2		H	6.9 ± 0.38	*nt*	**Opiate**, 5HT2a[h]	0.6	6.3	8.9	*nt*
[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
14		H	7.1 ± 0.22	50% inhib. at 10 μM	H3, Ca^{2+}[h]	1.1	6.0	9.3	32
15		F	7.0 ± 0.27	61% inhib. at 10 μM	Clean[h]	*[1.4]*	*[5.6]*	*[9.3]*	*nt*
16		H	7.1 ± 0.38	*nt*	Na^+, muscarinic, 5HT1a[h]	*[0.5]*	*[6.6]*	*[10.6]*	17

656 | *Med. Chem. Commun.*, 2011, **2**, 655–660 This journal is © The Royal Society of Chemistry 2011

Structure	Identifier	Previous Structure
Correct Structure 	**Cmpd A**	**ChEMBL Structure** 

Even after including in model building efforts (**RLM Clearance**), this compound was predicted as an outlier. Hence needed careful review.

Valade et al, *Med Chem Commun*, **2011**, *2*, 655

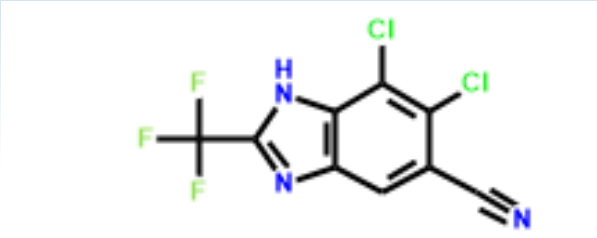
SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Outline

- What is data validation?
- Where do errors come from?
- **How to find them?**
- Why should we care about them?

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Activity/Toxicity Cliff (in database could be dubious)

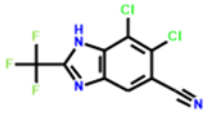
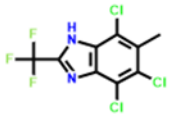
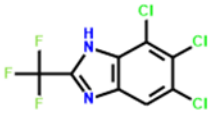
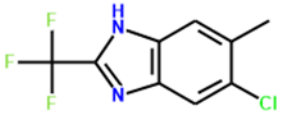
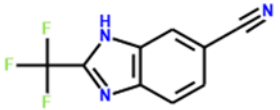
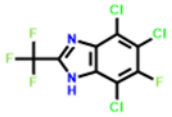
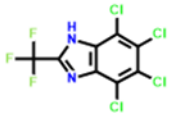
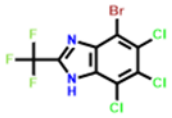


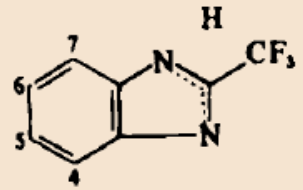
LD_{50} = 3955 mg/kg

31 most similar structures (sim > 0.8)

The LD_{50} range is 0.25-77 mg/kg.

https://chem.nlm.nih.gov/chemidplus/rn/startswith/89427-25-8

	Structure	Identifier	LD50_mgkg
1		89427-25-8	3955 **3995**
2		4228-99-3	18.286 **18.3**
3		2338-27-4	1.519 **1.52**
4		14863-40-2	55 **55**
5		3671-61-2	4.411 **4.41**
6		18225-94-0	2.386 **2.39**
7		2338-29-6	0.245 **0.25**
8		4228-93-7	4.122 **4.12**



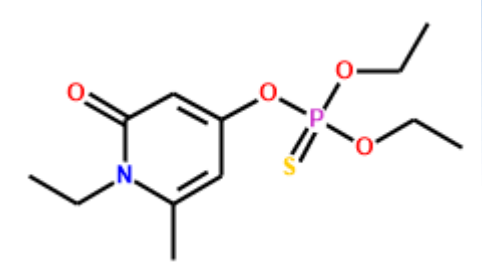
No.	Substituent	log LD_{50} (μmol kg^{-1})
1	—	2.17
2	5-Me	2.50
3	5-Et	2.11
4	5-*tert*-Butyl	2.16
5	5-F	1.91
6	4-Cl	1.94
86	4-NO_2;5,6-di-Cl	2.48
87	4,5-Di-Cl;6-NO_2	2.26
88	4,6-Di-Cl;5-NO_2	2.08
89	4,5-Di-Cl;6-CN	1.15
90	4,5,6,7-Tetra-F	1.11
91	4-F;5,6,7-tri-Cl	0.44
92	4,6-Di-F;5,7-di-Cl	0.76
93	4,6,7-Tri-Cl;5-F	0.89
94	4,6-Di-F;5,7-di-Br	1.28
95	4,5,7-Tri-Br;6-F	1.34

0.322 mg/kg

Adamson et al, *Pesticide Science*, **1984**, *15*, 31

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

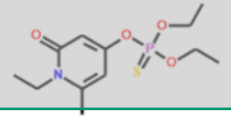
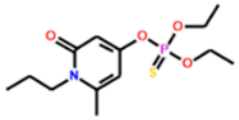
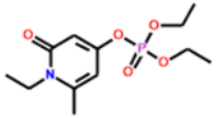
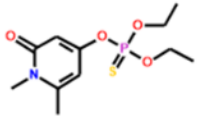
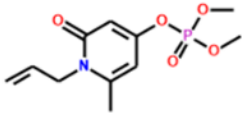
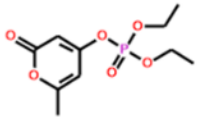
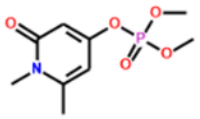
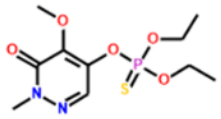
Activity/Toxicity Cliff (in database could be dubious)



LD50 = 7070 mg/kg

11 most similar structures

The LD50 range is 1.6-15 mg/kg.

Structure	Identifier	LD50_mgkg
	22787-58-2	7070 7070
	21327-31-1	9.75 9.75
	22787-59-3	1.62 1.62
	21409-78-9	5.97 5.97
	22620-72-0	2.46 2.46
	26662-09-9	3.34 3.34
	22787-53-7	2.91 2.91
	60244-60-2	2.5 2.5

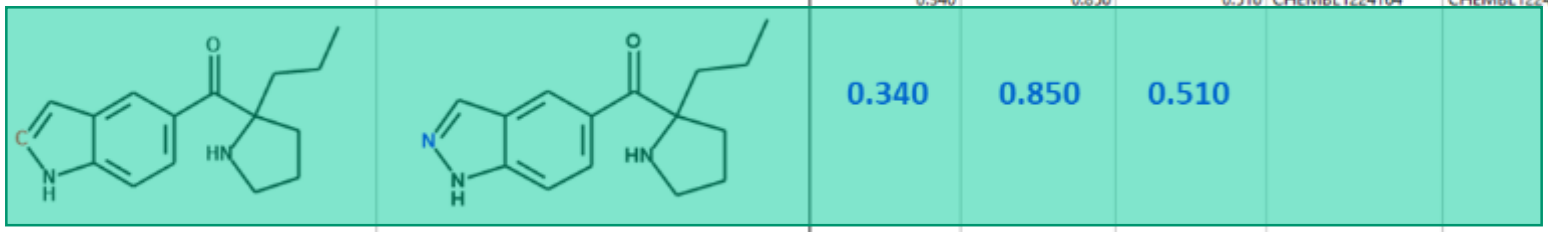
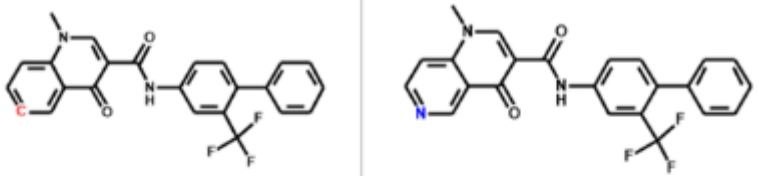
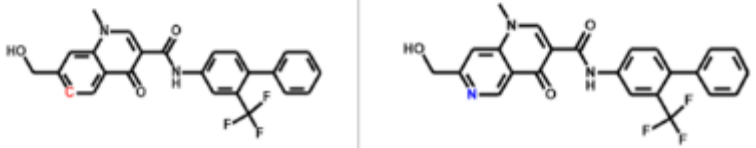
Our analysis suggested that expected LD_{50} could be in the range of 6-12 mg/kg

National Technical Information Service., OTS0571926

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Insights from Molecular Matched Pair Analysis

Rat Plasma protein binding data from a single research article.

	Structure 1	Structure 2	Rat_Fup 1	Rat_Fup 2	Change(Rat_Fup)	Identifier 1	Identifier 2
1			0.340	0.850	0.510	CHEMBL1224164	CHEMBL1224092
2			0.030	0.080	0.050	CHEMBL1938937	CHEMBL1938940
3			0.076	0.083	0.007	CHEMBL2037119	CHEMBL2037132
4			0.020	0.270	0.250	CHEMBL1378465	CHEMBL3393514

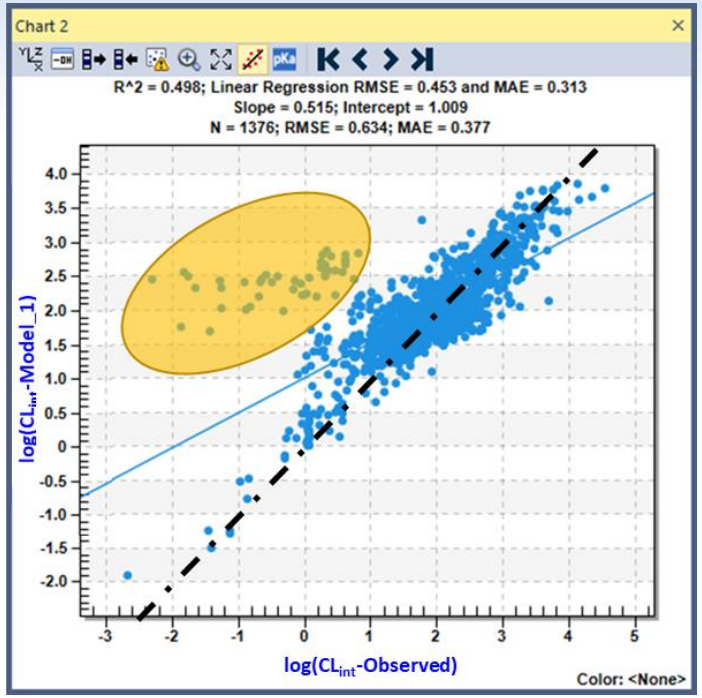
SimulationsPlus

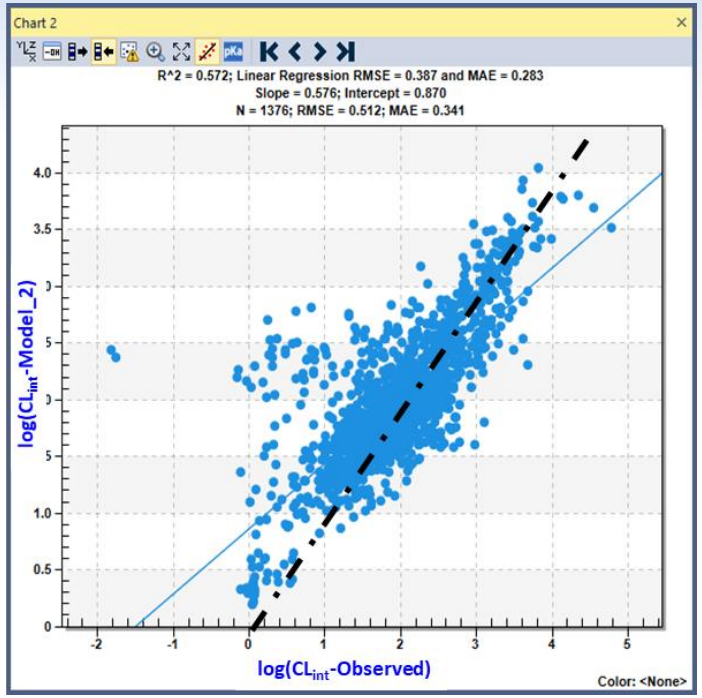
SCIENCE + SOFTWARE = SUCCESS

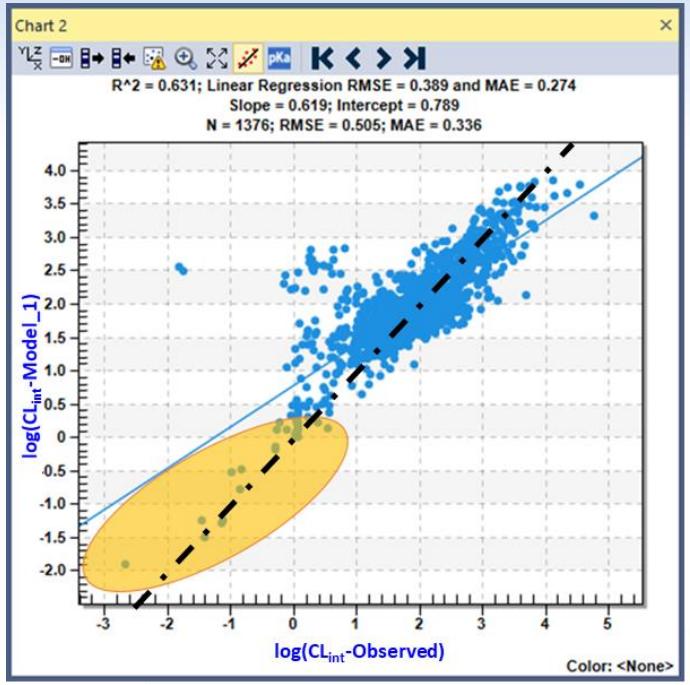
“Leave Group Out” in QSAR Model

helps to find the abnormalities in the dataset

- While building the Rat Liver Microsomal Clearance Model, a few cmpds needed correction of units & biological scaling
- Interim model suggested necessary corrections before including in the final dataset





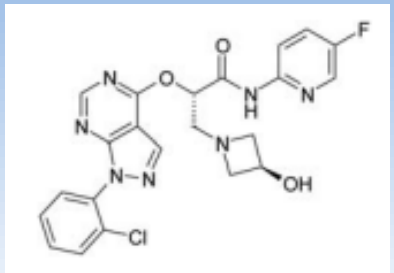


Reported units: µL/min/mg of protein
Actual units : mL/min per gm of liver

Unit discrepancies… 1000 fold

Knowledge is Knowing that Tomato is a Fruit…

…wisdom is not putting it in a fruit salad.



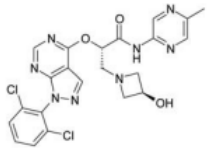
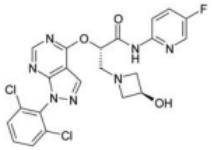
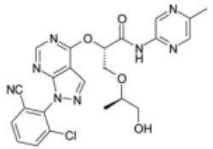
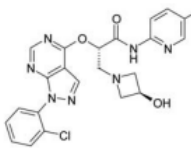
SAR Data | ADME Data | Metabolite Data | Toxicity Data | Clinical Data

- Activity Information

Assay Type	Common Name	Activity Mechanism	Entrez ID	Source	Activity Type	Activity
B	GLUCOKINASE	ACTIVATOR			pEC50	= 0.078
B	POTASSIUM CHANNEL KV11.1	INHIBITOR	3757	HUMAN	pIC50	= 36

Table 4 Representative compounds identified in the matrix exercise

Table 4 Representative compounds identified in the matrix exercise

	37/AZD9485	38	39/AZD3651	40
Structure				
GK EC_{50}/μM	0.096	0.049	0.078	0.078
log *D*	2.3	2.7	1.9	2.3
Solubility/μM	>750	80	>50	>750
MDCK P_{app} A to B/nm s^{-1}	35	73	78	98
hERG IC_{50}/μM	62	19	>100	36
Rat $Cl(Cl_u)$/mL min^{-1} kg^{-1}	25 (74)	42 (320)	18 (39)	31 (91)
Rat V_{ss}/L kg^{-1}	1.2	1.7	1.5	1.5
Rat *F*%	38	28	100	75
Dog $Cl(Cl_u)$/mL min^{-1} kg^{-1}	18 (60)	30 (210)	18 (49)	51 (170)
Dog V_{ss}/L kg^{-1}	1.1	1.1	0.9	1.5
Dog *F*%	87	32	43	12

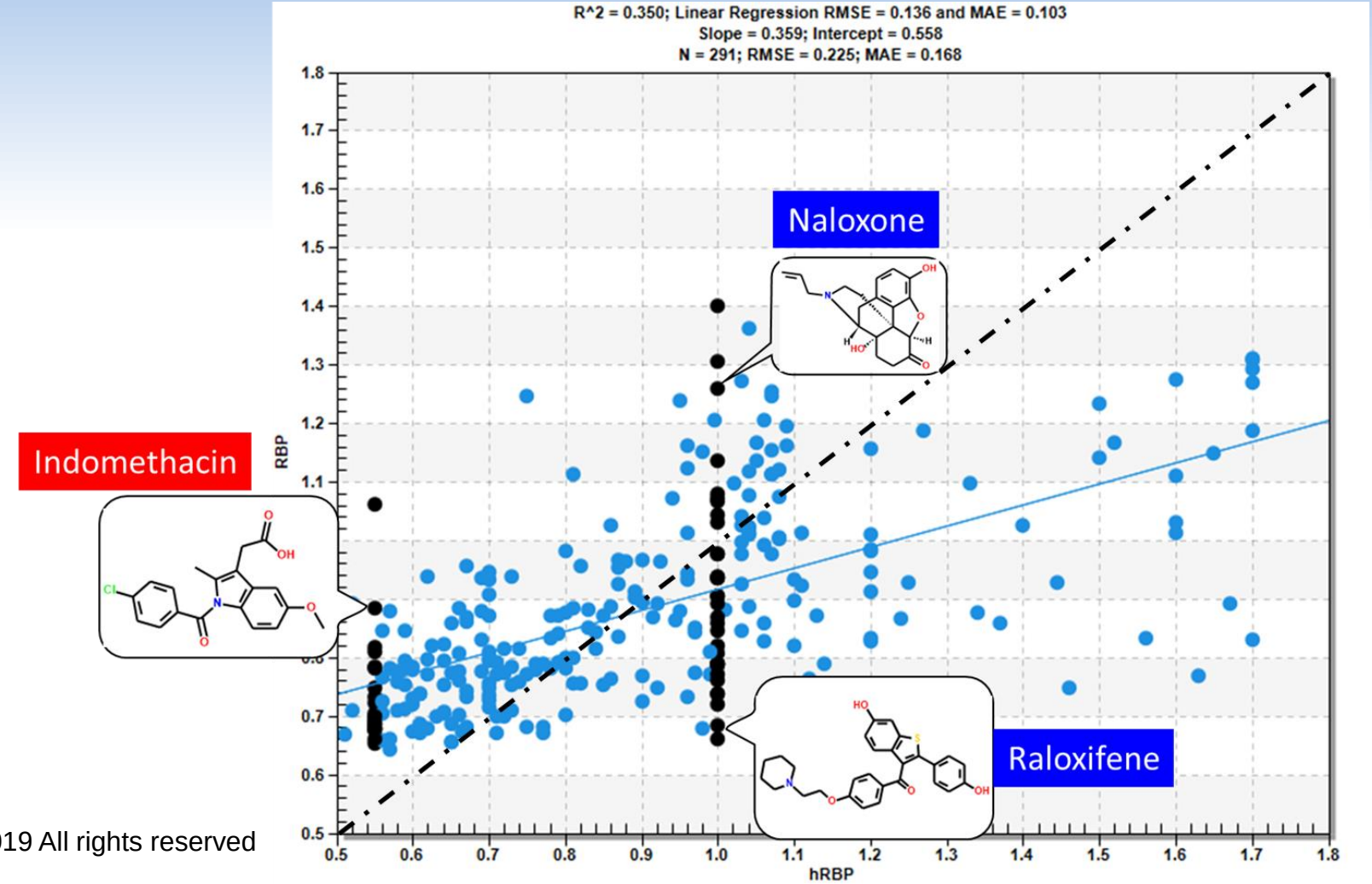
Many times, we just miss on removing a few data points that are wayyyyyyyyy (?) out of the **NORMAL RANGES**

Waring et al, *Med Chem Commun*, **2013**, *4*, 663

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

35 Different Chemotypes Have Identical RBP Values

For 35 compounds: Blood:Plasma Ratio = **1.00**
For 22 compounds: Blood:Plasma Ratio = **0.55**



Blood:Plasma Ratio : *"Mischief Managed"*

If no data is available, value of RBP is assumed

For Acidic compounds: **0.55**

For Basic compounds: **1.00**

Drug	In Vivo Parameters[a]			
	$CL^{i.v.}$	Renal CL	fu, p	R_B
	$ml \cdot min^{-1} \cdot kg^{-1}$			
Diclofenac	4.84	0.06	0.004	0.71[c]
Gemfibrozil	1.70	0.02	0.005[c]	0.75[c]
Mycophenolic acid	2.49	0.01	0.010	0.60
Naloxone	21.70	0.00	0.570	1.00[e]
Propofol	27.72[f]	0.00	0.015	0.88
Telmisartan	12.32	0.00	0.005	1.24[c]

[a] References are listed elsewhere (Supplemental Tables 2–6).
[b] $f_{m,\,UGT}$ values taken from in-house data were available.
[c] Measurement made in vitro.
[d] No suitable in vitro data were available so in vivo $f_{m,\,UGT}$ was used.
[e] No data available; value of 1 assumed for basic drugs.
[f] Blood clearance data.

TABLE 3

Mean intravenous and oral plasma clearance data, number of subjects, blood/plasma ratios

References for all the clinical data are listed in the supplemental data.

	Observed In Vivo Plasma Clearance[a]		R_B[b]
	Intravenous	Oral	
	$ml \cdot min^{-1} \cdot kg^{-1}$		
Quercetin	5.2 (0.6–12)	391 (1.55–62381)	1
Raloxifene	14.7[c]	735 (735–831)	1
Salbutamol	8.4 (7.6–9.3)	16.6 (12.6–22.9)	1
Troglitazone	2.5	12.8 (5.3–14.6)	0.55

N.A., no subject information available.
[a] Data are weighted mean (range). For salbutamol, weighted renal clearance was 4.8 and 4.3 $ml \cdot min^{-1} \cdot kg^{-1}$ and negligible for other drugs.
[b] Assumed to be 1 for basic drugs and 0.55 for acidic drugs.
[c] Calculated from oral clearance data (735 $ml \cdot min^{-1} \cdot kg^{-1}$) and reported bioavailability of 2% (Hochner-Celnikier,

TABLE 5

Pharmacokinetic parameters

Drug	CL_{tot}	R_b	CL_r	F	$F_a F_g$
	$ml \cdot min^{-1} \cdot kg^{-1}$		$ml \cdot min^{-1} \cdot kg^{-1}$		
Quercetin	11	1.0	0	0.001[a]	0.0
Raloxifene	14.7[b]	1.0	0	0.02	0.
Bazedoxifene	6.7[c]	0.55	0	0.06	0
Diclofenac	3.5	0.63	0	0.54	0.800
Tolfenamic acid	2.2	0.66	0	0.60	0.
Tolcapone	1.9	0.61	0	0.60	
Entacapone	12	0.59	0	0.25	1.0
Telmisartan	8.4	0.67	0	0.43	1.
Gemfibrozil	1.7	0.58	0	0.98	1.
Etodolac	0.66[d]	0.55	0	0.80	0.
Indomethacin	1.4	0.55[e]	0.21	1.0	1.00[f]

[a] Calculated from CL_{tot} (11 $ml \cdot min^{-1} \cdot kg^{-1}$) and oral clearance (8333 $ml \cdot min^{-1} \cdot kg^{-1}$) (Cubitt
[b] Calculated from oral clearance data (735 $ml \cdot min^{-1} \cdot kg^{-1}$) and reported bioavailability of 2%
[c] Calculated from oral clearance data (86.7 $ml \cdot min^{-1} \cdot kg^{-1}$) and reported bioavailability of
[d] Calculated from oral clearance data (0.82 $ml \cdot min^{-1} \cdot kg^{-1}$) and reported bioavailabi
(http://www.accessdata.fda.gov/drugsatfda_docs/label/2006/018922s022,020584s009lbl.pdf).
[e] Calculated $F_a F_g$ was < 0, because $CL_{tot}/R_b/Q_h$ was > 1. Therefore, CL_{tot}/R_b was assumed to 90%
[f] When calculated $F_a F_g$ was > 1, it was treated as 1.0.
[g] Assumed to be 0.55 because indomethacin and etodolac are acidic drugs (Cubitt et al., 2011).

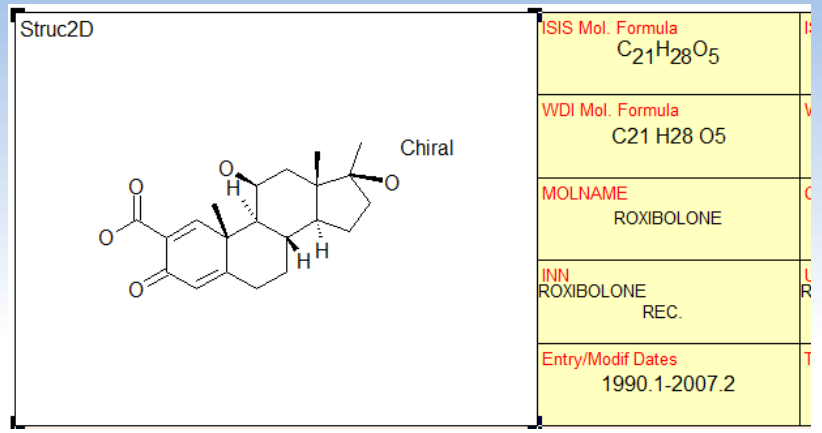
Gill et al, *Drug Metab Dispos*, **2012**, *40*, 825

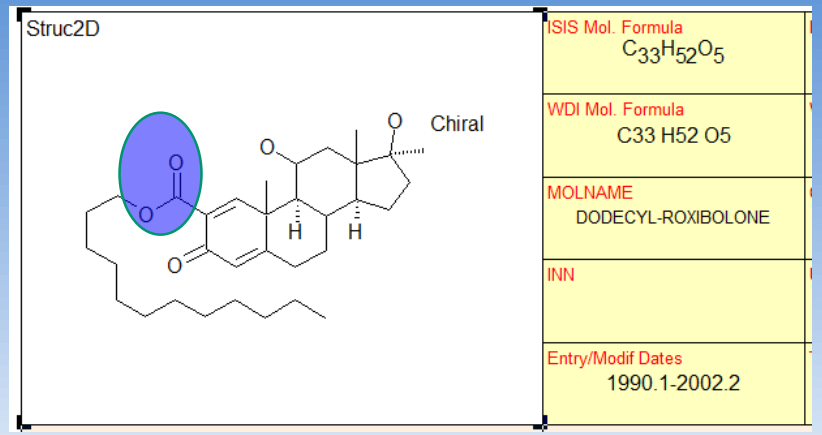
Cubitt et al, *Drug Metab Dispos*, **2011**, *39*, 864

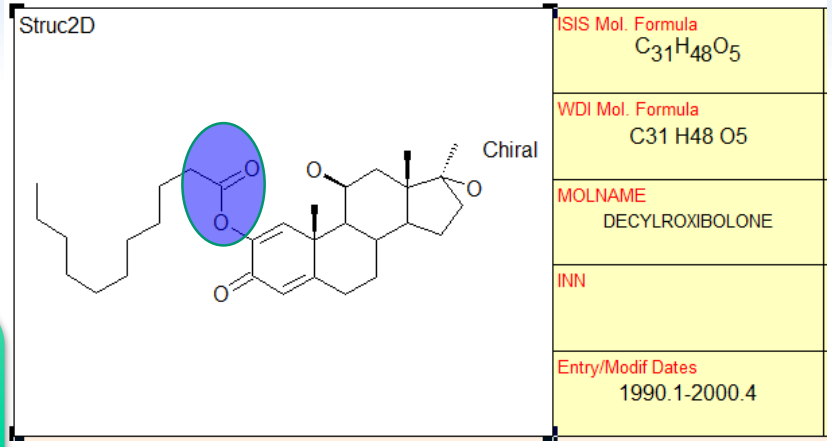
Nakamori et al, *Drug Metab Dispos*, **2012**, *40*, 1771

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

User Experience







Prodrugs

- The ester in the "decyl" ester is reversed.
  - While testing esterase transforms in ADMET Predictor, we found a prodrug hydrolysis product with 11 carbons

Mammals have problems metabolizing odd-carbon long chain fatty acids, so they are not very common when it comes to prepare the prodrugs.

Courtesy: Bob Clark's Diaries

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Outline

- What is data validation?
- Where do errors come from?
- How to find them?
- Why should we care about them?

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

A lie repeated a thousand times becomes truth…

Daily Mail

NO MORE SURRENDERS

LONDON SURGEON'S PHOTO OF THE MONSTER

Monster Yards from Lochside: See Enlargement Inside

Does Monster Really Exist?

In 1993, Duke Wetherell, admitted he'd made the "monster" out of some plastic and a clockwork, tinplate, toy submarine.
(1934 – 1993)

Monday, January 11, 1999

Bigfoot

New claim

Ray Wallace died at 84. His family then revealed a secret: The whole thing was a hoax. Wallace had made the oversized footprints with a set of carved wooden feet.
(1957 – 1999)

BBC | Sign in | News | Sport | Weather | Shop | Earth | Travel | More | Search

newsbeat

Latest | Popular | Topics

2017 has been a 'record year' for sightings of the Loch Ness monster

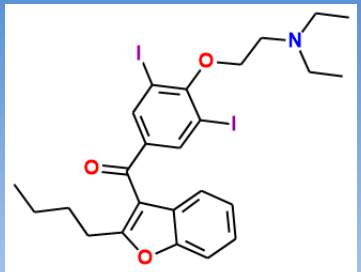
YOU WHAT? | 15 Nov 2017

FOX26NEWS

Bigfoot sighting reported in California

by Alexandra Lehnert, KMPH | Thursday, October 12th 2017

Amiodarone Vs Amiodarone HCl



CID 2157

Amiodarone

Solubility

Low

- from DrugBank

Soluble @ 25 deg C (g/100 ml): chloroform 44.51; methylene chloride 19.20; methanol 9.98; ethanol 1.28; benzene 0.65; tetrahydrofuran 0.60; acetonitrile 0.32; 1-octanol 0.30; ether 0.17; 1-propanol 0.13; hexane 0.03; petroleum ether 0.001; sparingly soluble in iso- propanol; slightly soluble in acetone, dioxane, and carbon tetrachloride

O'Neil, M.J. (ed.). The Merck Index - An Encyclopedia of Chemicals, Drugs, and Biologicals. 13th Edition, Whitehouse Station, NJ: Merck and Co., Inc., 2001., p. 85

- from HSDB

In water, 700 mg/l @ 25 deg C

O'Neil, M.J. (ed.). The Merck Index - An Encyclopedia of Chemicals, Drugs, and Biologicals. 13th Edition, Whitehouse Station, NJ: Merck and Co., Inc., 2001., p. 85

- from HSDB

CID 441325

Amiodarone hydrochloride

Solubility

Soluble @ 25 deg C (g/100 ml): chloroform 44.51; methylene chloride 19.20; methanol 9.98; ethanol 1.28; benzene 0.65; tetrahydrofuran 0.60; acetonitrile 0.32; 1-octanol 0.30; ether 0.17; 1-propanol 0.13; hexane 0.03; petroleum ether 0.001; sparingly soluble in iso- propanol; slightly soluble in acetone, dioxane, and carbon tetrachloride

O'Neil, M.J. (ed.). The Merck Index - An Encyclopedia of Chemicals, Drugs, and Biologicals. 13th Edition, Whitehouse Station, NJ: Merck and Co., Inc., 2001., p. 85

- from HSDB

In water, 700 mg/l @ 25 deg C

O'Neil, M.J. (ed.). The Merck Index - An Encyclopedia of Chemicals, Drugs, and Biologicals. 13th Edition, Whitehouse Station, NJ: Merck and Co., Inc., 2001., p. 85

- from HSDB

While building solubility models **ADMET Modeler** consistently finds it as an outlier

Its nominal solubility of 0.7 mg/mL is predicted to be much less by **ADMET Predictor**

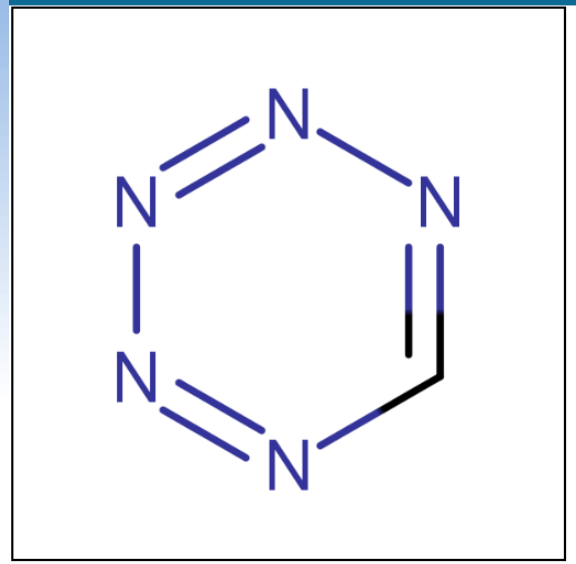
https://pubchem.ncbi.nlm.nih.gov/compound/2157
https://pubchem.ncbi.nlm.nih.gov/compound/441325

Courtesy: Bob Clark's Diaries

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Pentazine…

United States Environmental Protection Agency



Wikipedia

Pentazine is a hypothetical compound that consists of a six-membered aromatic ring containing five nitrogen atoms with the molecular formula CHN_5. The name *pentazine* is used in the nomenclature of derivatives of this compound.
...
Read more

Intrinsic Properties

Molecular Formula: CHN_5 Mol File
Find All Chemicals
Average Mass: 83.054 g/mol
Isotope Mass Distribution
Monoisotopic Mass: 83.023195 g/mol

PubChem Pentazine (Compound) CONTENTS

2.3.1 CAS

290-97-1

from ChemIDplus; EPA DSSTox

PubChem Pentazine (Compound)

5 Chemical Vendors

Showing 1 Substance per Vendor View All View in Entrez Download

Chem-Space.com Database
PubChem SID: 343047693 Purchasable Chemical: CSC010257012

ChemTik
PubChem SID: 162499770 Purchasable Chemical: CTK1A5264

Ambinter
PubChem SID: 368788451 Purchasable Chemical: Amb25295300

from PubChem

… is a nonexistent compound that has a DSSTOx record.
… got into the literature for computational work trying to rationalize the fact that it does not exist
… got a CAS Number
… is (surprisingly) available to purchase from chemical vendors

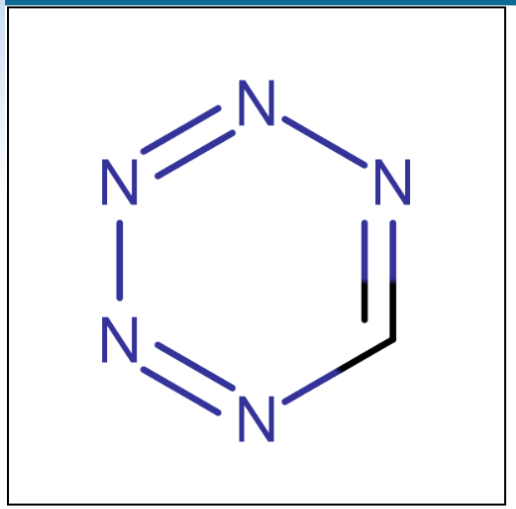
Once that happened, it became "*virtually real*" regardless.

Courtesy: Marvin Waldman's Diaries

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Pentazine…..Possible Hazards

United States
Environmental Protection
Agency

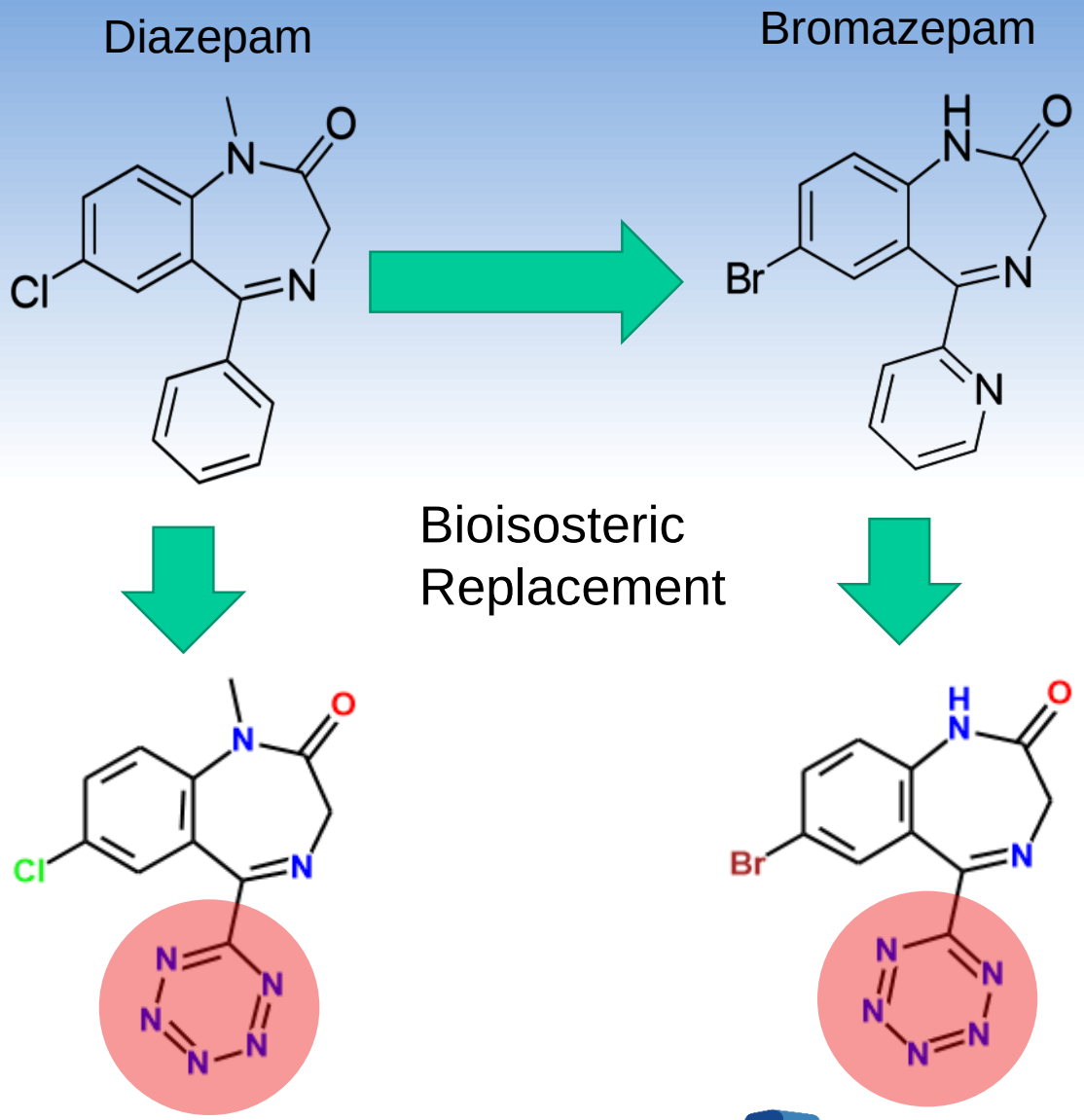


Wikipedia

Pentazine is a hypothetical compound that consists of a six-membered aromatic ring containing five nitrogen atoms with the molecular formula CHN_5. The name *pentazine* is used in the nomenclature of derivatives of this compound.
...
Read more

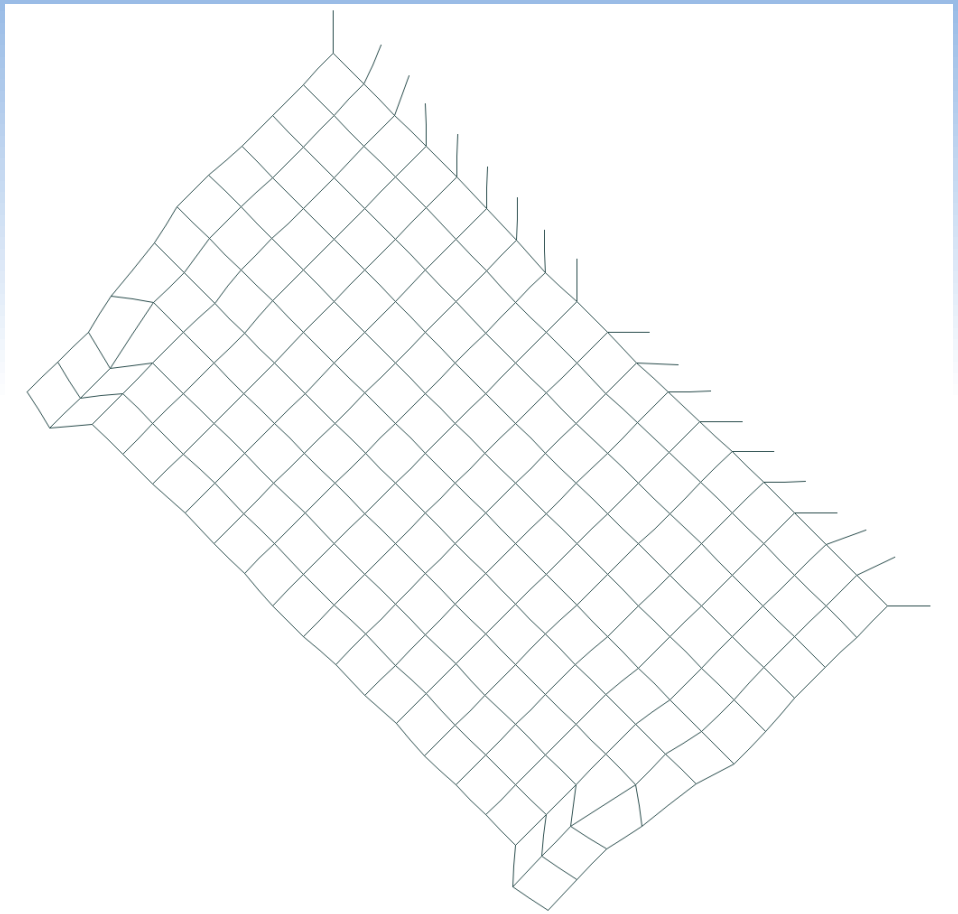
Intrinsic Properties

Molecular Formula: CHN_5 Mol File
Find All Chemicals
Average Mass: 83.054 g/mol
Isotope Mass Distribution
Monoisotopic Mass: 83.023195 g/mol



SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

CID 20681682



Molecular Formula:	$C_{228}H_{98}$
Molecular Weight:	2837.292 g/mol

US6281371

(12) **United States Patent**
Klösel et al.

(54) **LIPOPOLYAMINES, AND THE PREPARATION AND USE THEREOF**

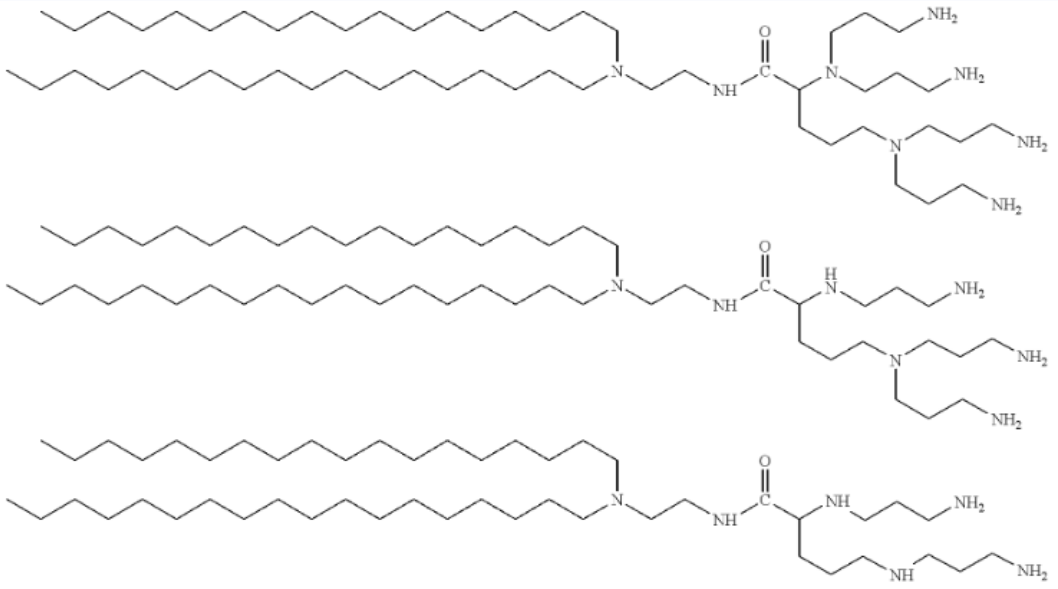
(75) Inventors: **Roland Klösel**; **Stephan König**, both of München (DE)

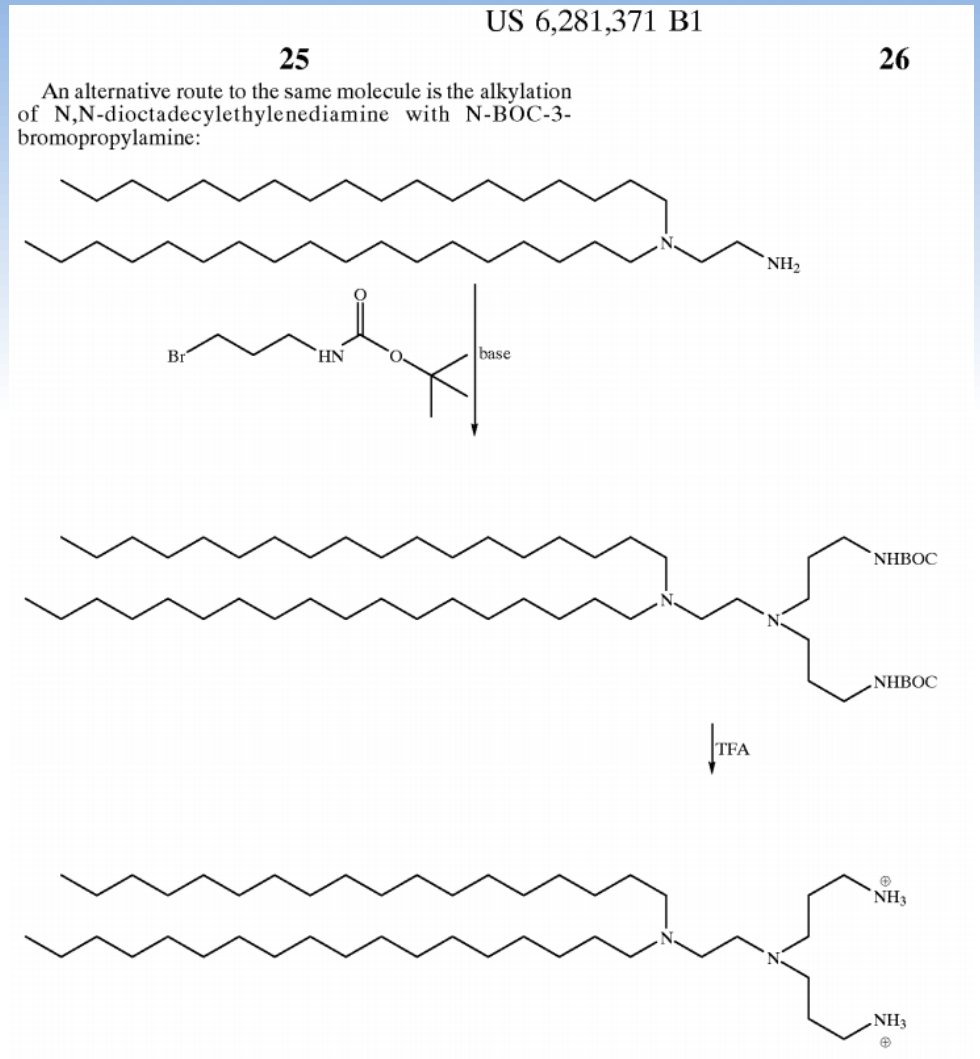
(73) Assignee: **Biontex Laboratories GmbH**, Munich (DE)

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

CID 20681682… story continues

All compounds in the patent are various lipopolyamines and substances (raw materials) to synthesize them

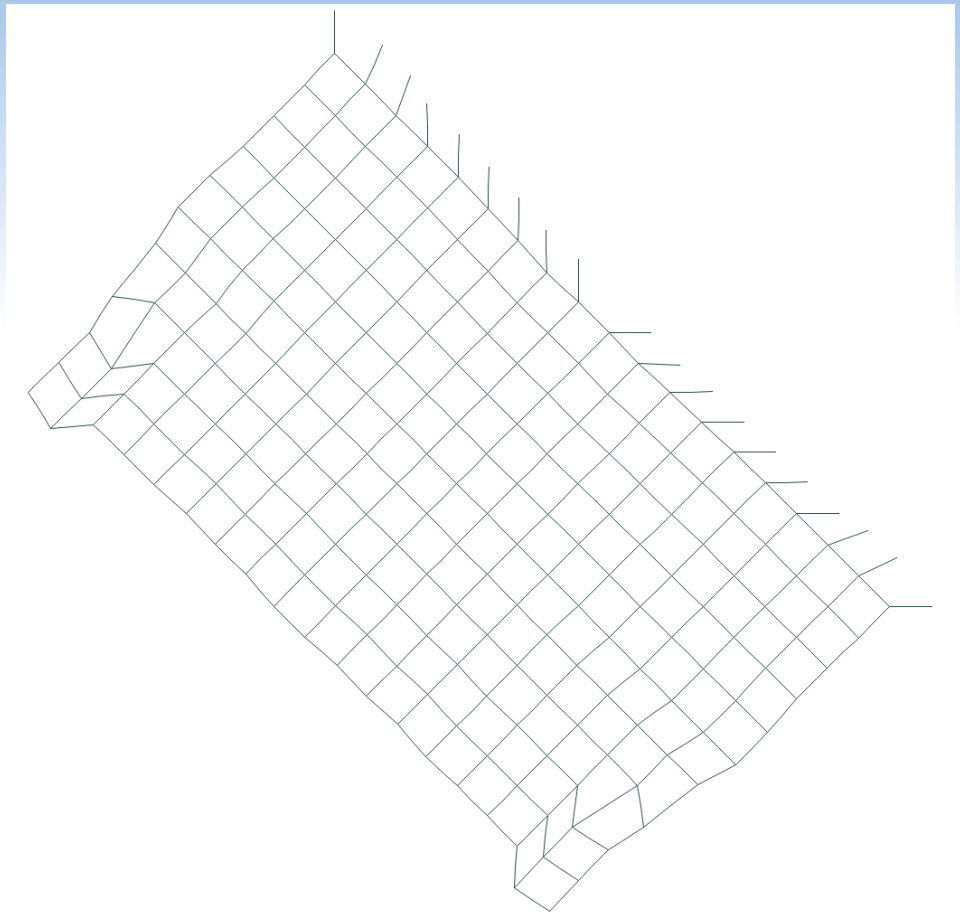


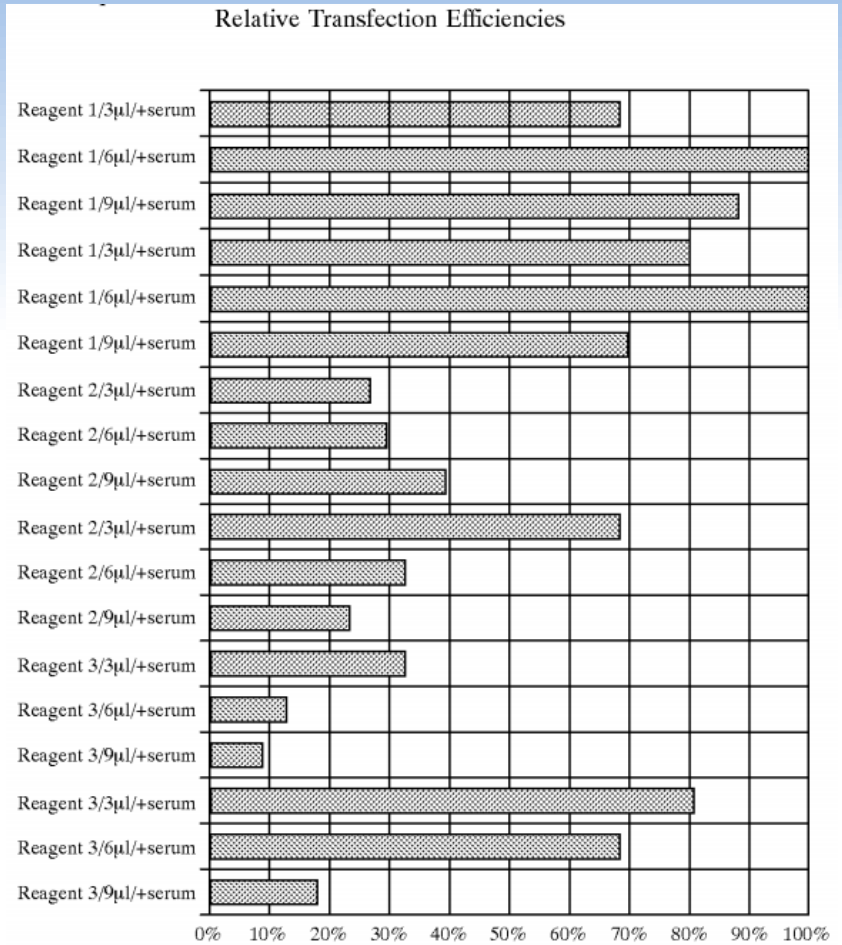


SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

The Wages of A(B)CD's

The only graphics that looks similar to CID 20681682 is a bar-chart depicting comparison of transfection efficiencies

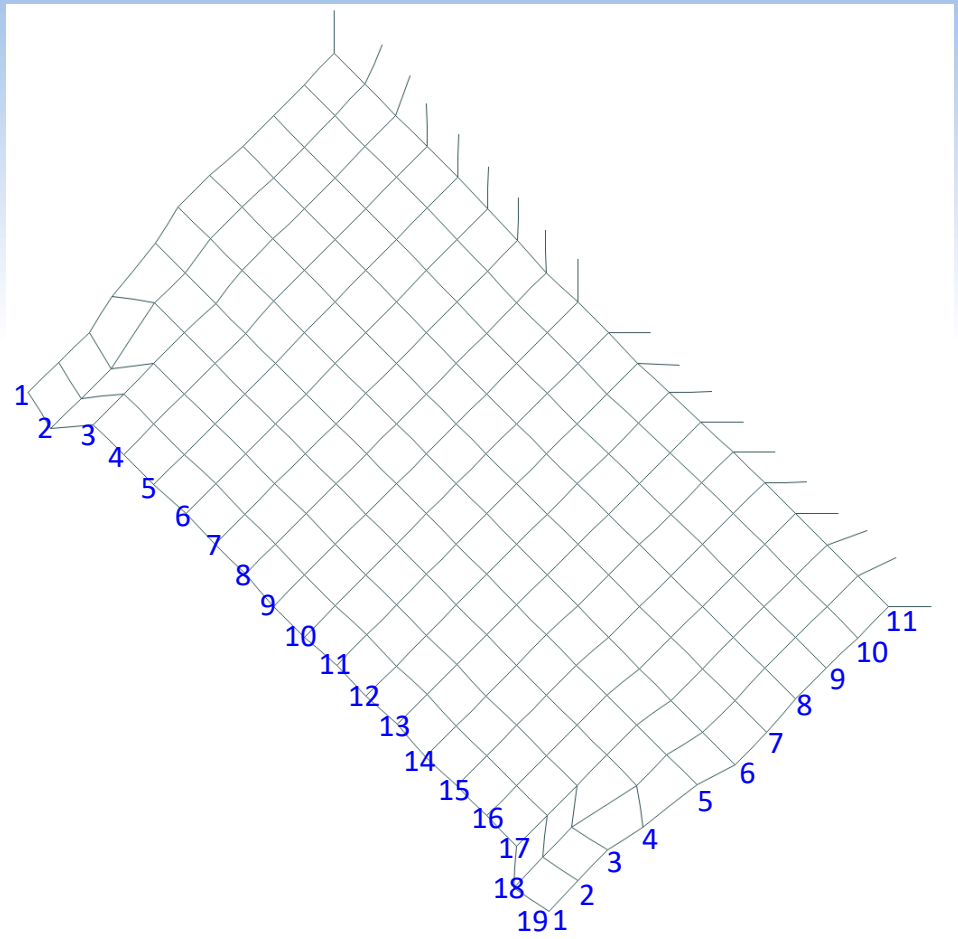


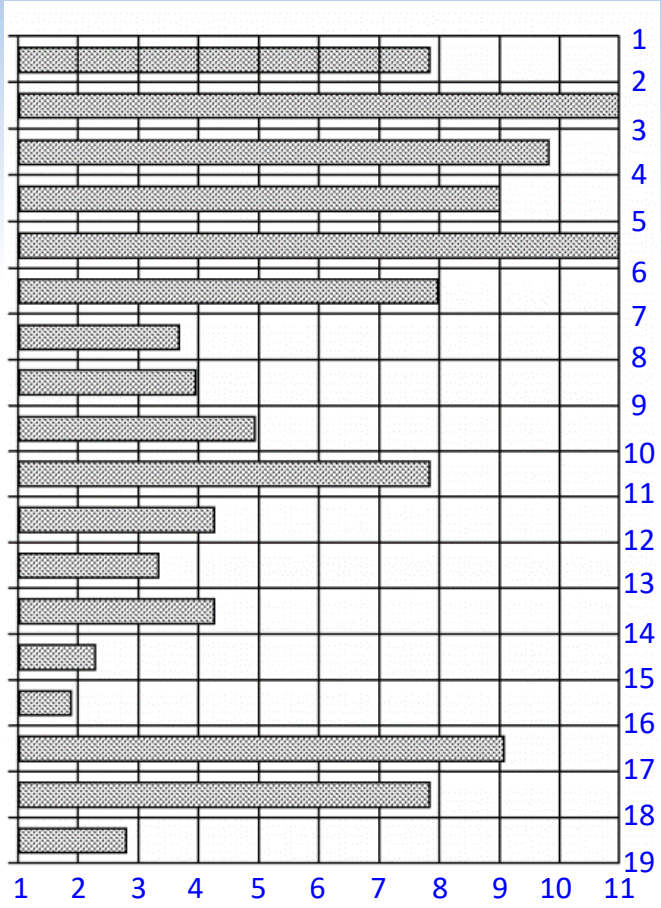


SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

The Wages of A(B)CD's

Assumption is that **CID-20681682** is a result of *A(B)CD* as the bar chart and the cmpd have identical grid-dimensions including extra "spikes"





SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Validation is Necessary:
Be it News or Activity Data

Daily Mail 1934

New York Times 1988

THE NEW YORK TIMES, TUESDAY, MARCH 22, 1988

Are Scientists a Threat to Rare 'Fossil Fish'?

Continued From Page C1

German scientists had used a small submarine to film coelacanths in their natural habitat off the Comoros islands. The German group, led by Dr. Hans Fricke of the Max Planck Institute for Comparative Physiology in Seewiesen, succeeded for the first time in photographing the peculiar movements and feeding habits of the five-foot fish at the bottom of the ocean. Coelacanths sometimes perform headstands or swim upside down as they loll along the ocean floor.

Subsequently, the New York Explorers Club, the New York Aquarium and a consortium of academic institutions organized an effort to capture and transport a live coelacanth to an aquarium, or, failing that, to acquire dead specimens for dissection. The organizers published an invitation to volunteers willing to pay $4,000 in costs to join a coelacanth expedition.

A nine-member delegation, including two paying volunteers, went to the Comoros last November in search of coelacanths. The group caught no live animals but acquired two frozen specimens from sources that scientists declined to identify. A weeklong post-mortem examination of the two fish was conducted in January at the Virginia Institute of Marine Science by about 30 scientists from 10 universities.

Clarifying Role in Evolution

From computed tomography (CT) scans, microphotographs of tissue cells and analysis of DNA collected in the post-mortem, the team hopes to clarify the role the coelacanth may have played in the evolution of land animals.

Louis E. Garibaldi, acting director of the New York Aquarium and a leader of the campaign to capture a coelacanth, said in an interview that his group had established a network of contacts among fishermen and others in the Comoros, so that if a coelacanth was captured alive it could be protected and swiftly moved to an aquarium.

Looking for Adventure, Excitement, Discovery? Join Expedition Seeking Ancient Coelacanth

The International Angler

Headline of notice recruiting volunteers willing to pay to join an expedition to capture a coelacanth.

Hans Fricke, who opposes such expeditions, and an assistant, photographing dead coelacanths.

catches. For one thing, they're using much heavier lines than in the past, to hold the big coelacanths when they hook them."

Dr. Fricke said he easily could have captured coelacanths and maintained them under pressure in his own sub-

Breeding coelacanths in captivity might help save the species, some assert

canth's domain in June, supported in part by the National Geographic Society. Meanwhile, he is pursuing another line of research based on clues found in 19th century Spanish churches.

"Some of these churches contain

Press & Sun Bulletin 1999

Monday, January 11, 1999

LIVIN

Bigfoot or big lie?

New claims against famous film se

BOTHELL, Wash. (AP) — In the hearts and minds of true believers, Bigfoot's existence has long been confirmed in a single minute of jerky, grainy footage of a startled sasquatch retreating into the upper California woods.

But two enthusiasts of the legendary hairy giant are alleging four magnified frames of the 59-year-old footage show traces of a bell-shaped fastener at Bigfoot's waist. They say the creature in the so-called Patterson-Gimlin Film can finally be proved as a man or a monkey

said Cliff Crook, pot tracker who sasquatch more arbiters of hoax artifical, non-ape or a Bigfoot?" totally showing a along a streambed, ger Patterson and ct. 20, 1967. It has nd independent many shaped as man-beast, has evidence of its very

of really detract Ray Crowe, pres- ana Bigfoot Soci- Ore. The image image "has a fluid eal creature of

portant because overcomparable all alltale footprints ade by Patterson totally filmed the across a sandbar National Forest. footage, experts gold standard for The washed away assertion on com- ents performed by Bigfoot buff from the Columbia, who covered an other- year in 1995 while

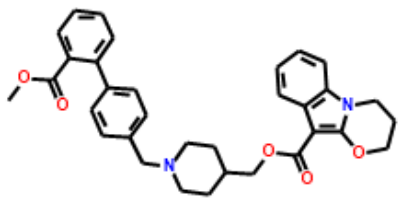
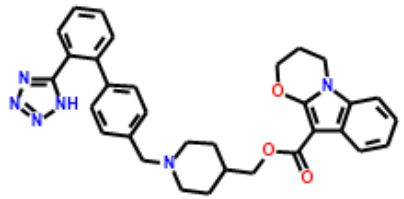
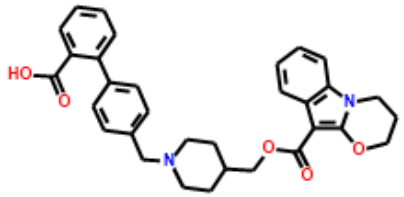
ASSOCIATED PRESS

Photographers Roger Patterson and Bob Gimlin made this image Oct. 20, 1967 purportedly showing a female Bigfoot. Now, four magnified frames of the footage show traces of a bell-shaped fastener at Bigfoot's waist, and after decades of doubt, some enthusiasts say the creature is a fake.

But always remember:
There really are (some)
black swans out there...

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Black Swans Do Exist: Activity/Toxicity Cliff

	Structure	Identifier	pIC50	IC50
1		CHEMBL2440459	6.883	131.000
2		CHEMBL2440456	4.561	27500.000
3		CHEMBL2440460	4.323	47500.000

- Not all Activity cliffs are **Mistakes/Errors**
- Some <u>Activity Cliffs</u> are Real. (hERG IC_{50} data)
- VALIDATE/VERIFY what you see is real and not an artifact or human error

Brudeli et al, *Bioorg Med Chem*, **2013**, *21*, 7134

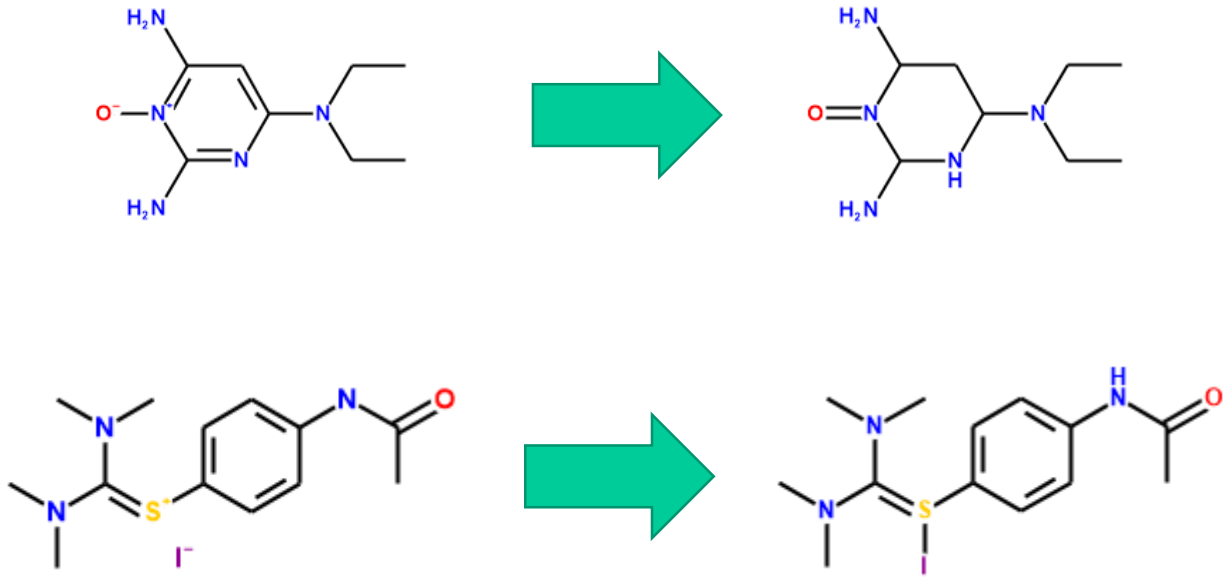
SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Can We Automate the Data Curation Process?

SAR AND QSAR IN ENVIRONMENTAL RESEARCH, 2016
VOL. 27, NO. 11, 911–937
http://dx.doi.org/10.1080/1062936X.2016.1253611

Taylor & Francis
Taylor & Francis Group

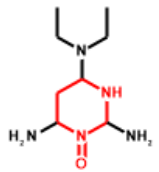
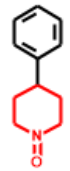
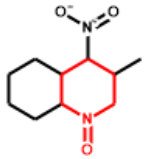
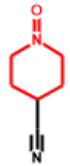
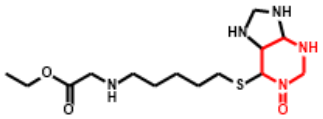
An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling[$]



Courtesy: Marvin Waldman's Diaries

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

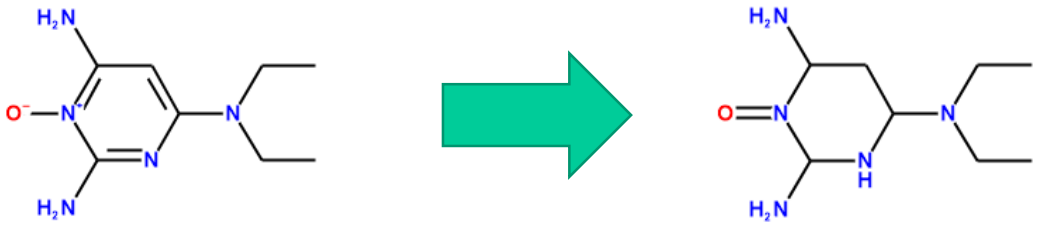
Can We Automate the Data Curation Process?

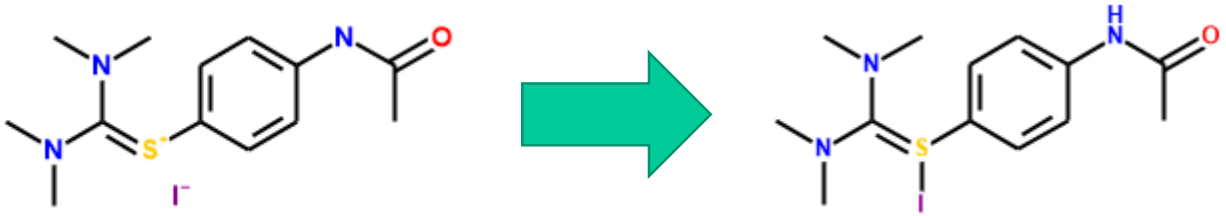
Structure	Identifier
	DTXSID90879247
	DTXSID9061548
	DTXSID8074944
	DTXSID3065827
	DTXSID00877618

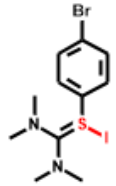
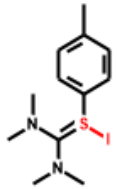
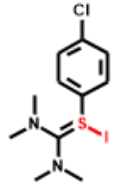
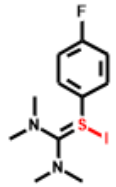
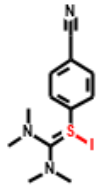
SAR AND QSAR IN ENVIRONMENTAL RESEARCH, 2016
VOL. 27, NO. 11, 911–937
http://dx.doi.org/10.1080/1062936X.2016.1253611

Taylor & Francis
Taylor & Francis Group

An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling[$]





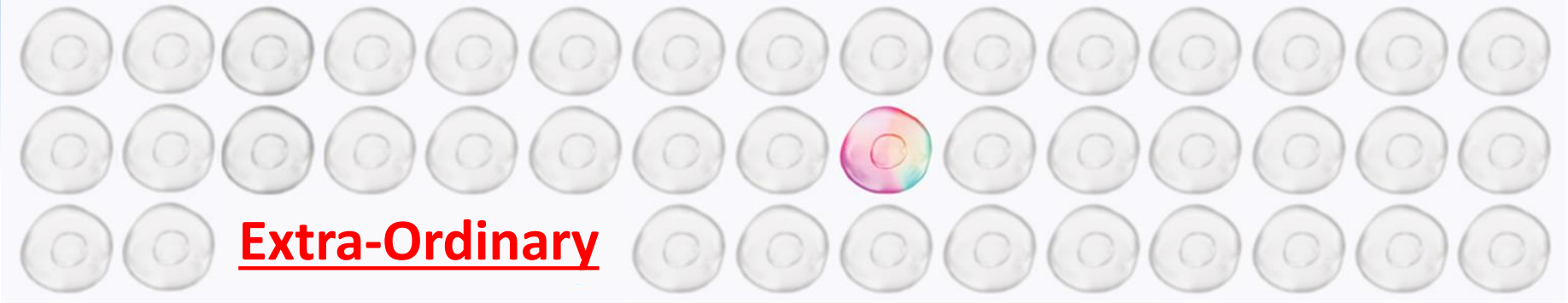
Structure	Identifier
	DTXSID20876522
	DTXSID40876963
	DTXSID60876526
	DTXSID30876533
	DTXSID40876862

Courtesy: Marvin Waldman's Diaries

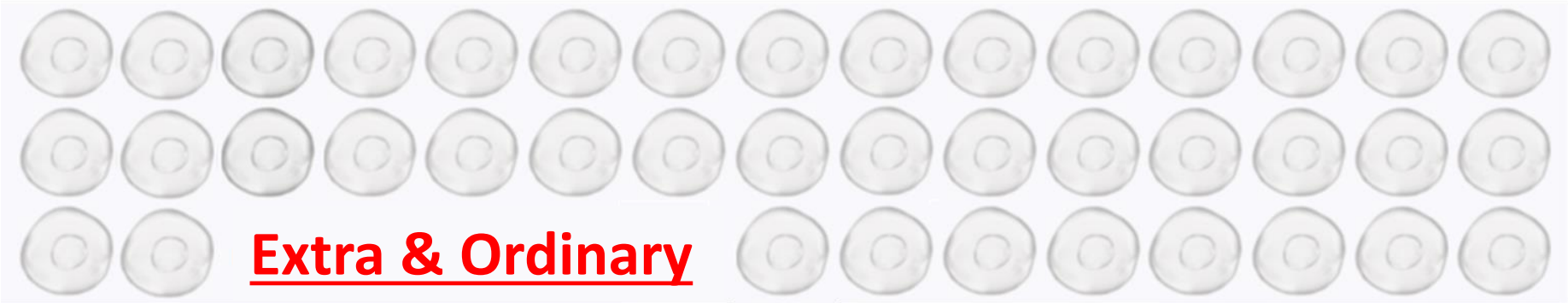
“To Validate or Not to Validate” That is The Question

A model can never be better than the data used to build it.

Good and correct data helps to build **Extra-Ordinary** model NOT just **Extra & Ordinary** Model



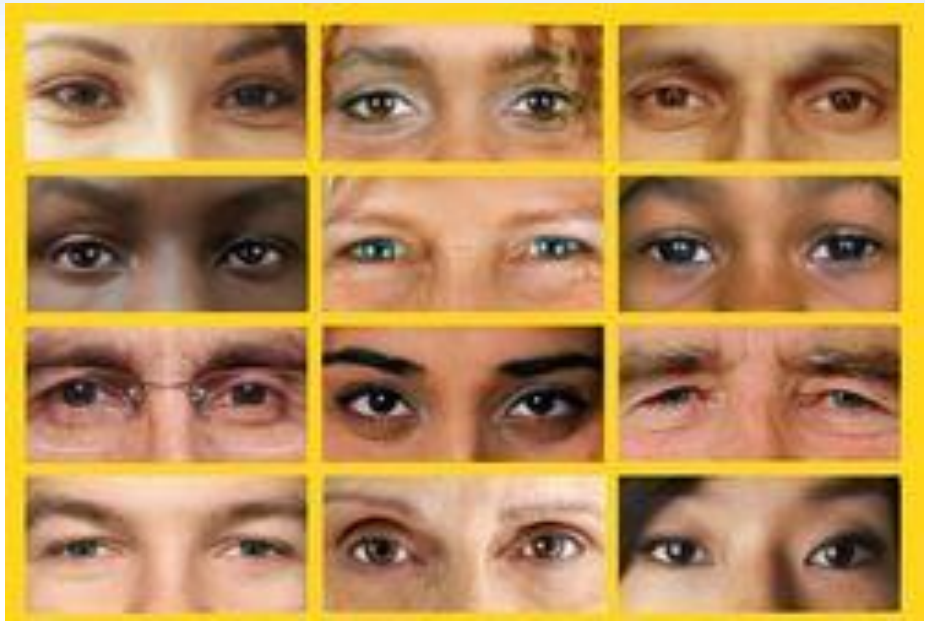
OR



SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

Conclusion

- Poor-quality data is enemy number one to the effective application of machine learning
- One needs to be vigilant while using any bioactivity databases or compilations
- Watch out for A(B)CD's
- Automation is necessary but it could be dangerous
- What is desired?
  - The Extra-Ordinary OR Extra & Ordinary ???
- If you see something, say something
  - Pears for your heirs



SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

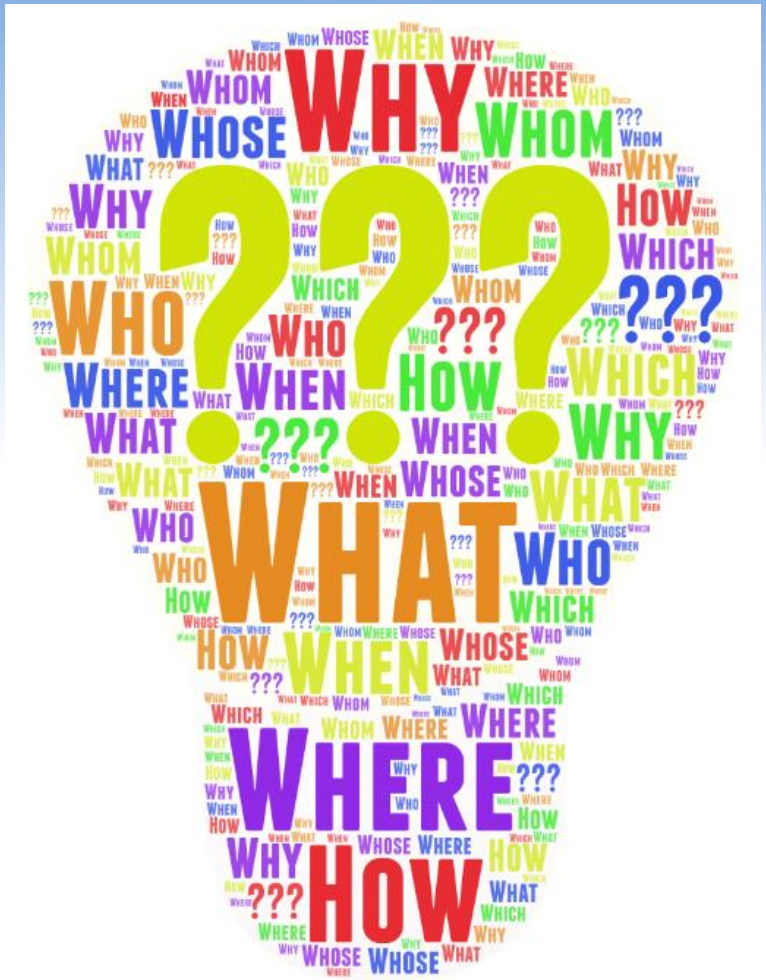
My thanks to:

Simulations Plus Team

- Bob Clark
- Marvin Waldman
- Mike Bolger
- Robert Fraczkiewicz
- David Miller
- Michael Lawless
- Eric Jamois
- John DiBella

OpenForce Field Initiative Team

- John Chodera
- David Mobley
- Michael R Shirts
- Michael Gilson
- Karmen Čondić-Jurkić



THANK YOU!!!

SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS