Model-Informed Drug Development

MIDD+

2021 Virtual Conference

Untold Stories of Data Curation

MDCK Project : Data Curation & Modeling

Phyo Phyo Kyaw Zin, Pankaj Daga, Michael Lawless, Bob Clark

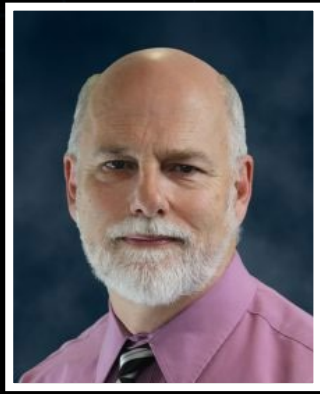
SimulationsPlus

Cognigen | DILIsym Services | Lixoft

Acknowledgement

Pankaj Daga

Michael Lawless



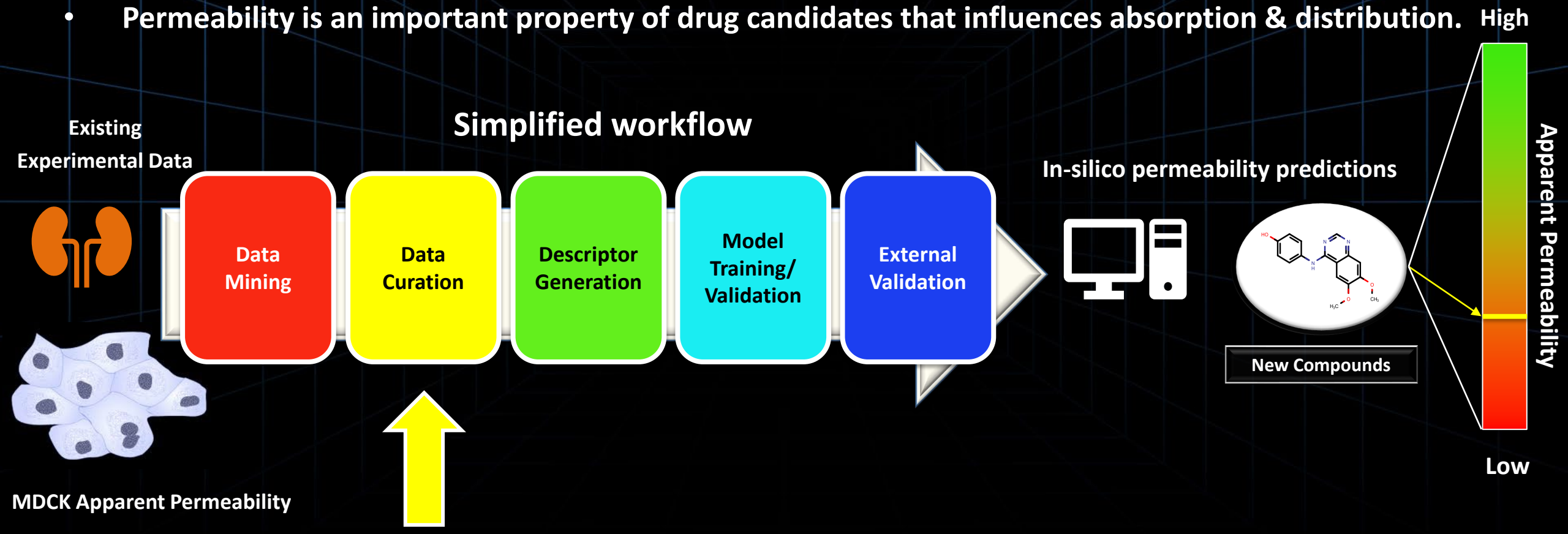
Robert Clark

Marvin Waldman

David Miller

MDCK (Madin-Darby Canine Kidney)

- MDCK is a popular mammalian cell line used to measure apparent permeability.
- Permeability is an important property of drug candidates that influences absorption & distribution.



MDCK Data Curation Workflow

Original Set
(13k entries)

Data mining from databases & literature ---------- *ChEMBL, GOSTAR, Litera*

Waldman, M., Fraczkiewicz, R., & Clark, R. D. (2015). Tales from the war on error: The art and science of curating QSAR data. *Journal of Computer-Aided Molecular Design*, *29*(9), 897–910.

Automation + human efforts

Cell line categorization ----------

Distinguishing experimental & predicted values

Fourches, D., Muratov, E., & Tropsha, A. (2011). Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research.

A fundamental assumption of any cheminformatics study is the correctness of the input data generated by experimental scientists and available in various datasets. Nevertheless, a recent study[6] showed that on average there are two errors per each medicinal chemistry publication with an overall error rate for compounds indexed in the WOMBAT database[7] as high as 8%. In another recent study[8], the authors investigated several public and commercial databases to calculate their error rates: the latter were ranging from 0.1 to 3.4% depending on the database.

An enormous body of data is now available for use in modeling quantitative structure/activity relationships (QSARs) and modern cheminformatics tools make it more accessible than ever before. Fortunately, most of the data are accurate, which is a testament to the skill and determination of those who produced it. Unfortunately, the fraction that is not accurate is uncomfortably large—estimated at up to 10 % overall [1, 2]. This is not surprising, given that access to the data is primarily through secondary sources—large public and commercial compilations—and

Data Analysis +
Automated Scripts +
Duplicate Assessment +
Human Intelligence +
Manual Assessment

Chemical Data Curation ---------- *Mistake Diagnosis and Correction*

Mistake Detection
(Papp Values, Cell-lines, Structures, Units) ----------

Manual Inspection ----------

Curated Set
(891 entries)

Mistakes Found During MDCK Data Curation Process

10.1021/acs.jmedchem.9b01411

After discarding unverifiable entries, ~270 for ~30% of the f

Mistakes are abundantly present in the databa l curation.

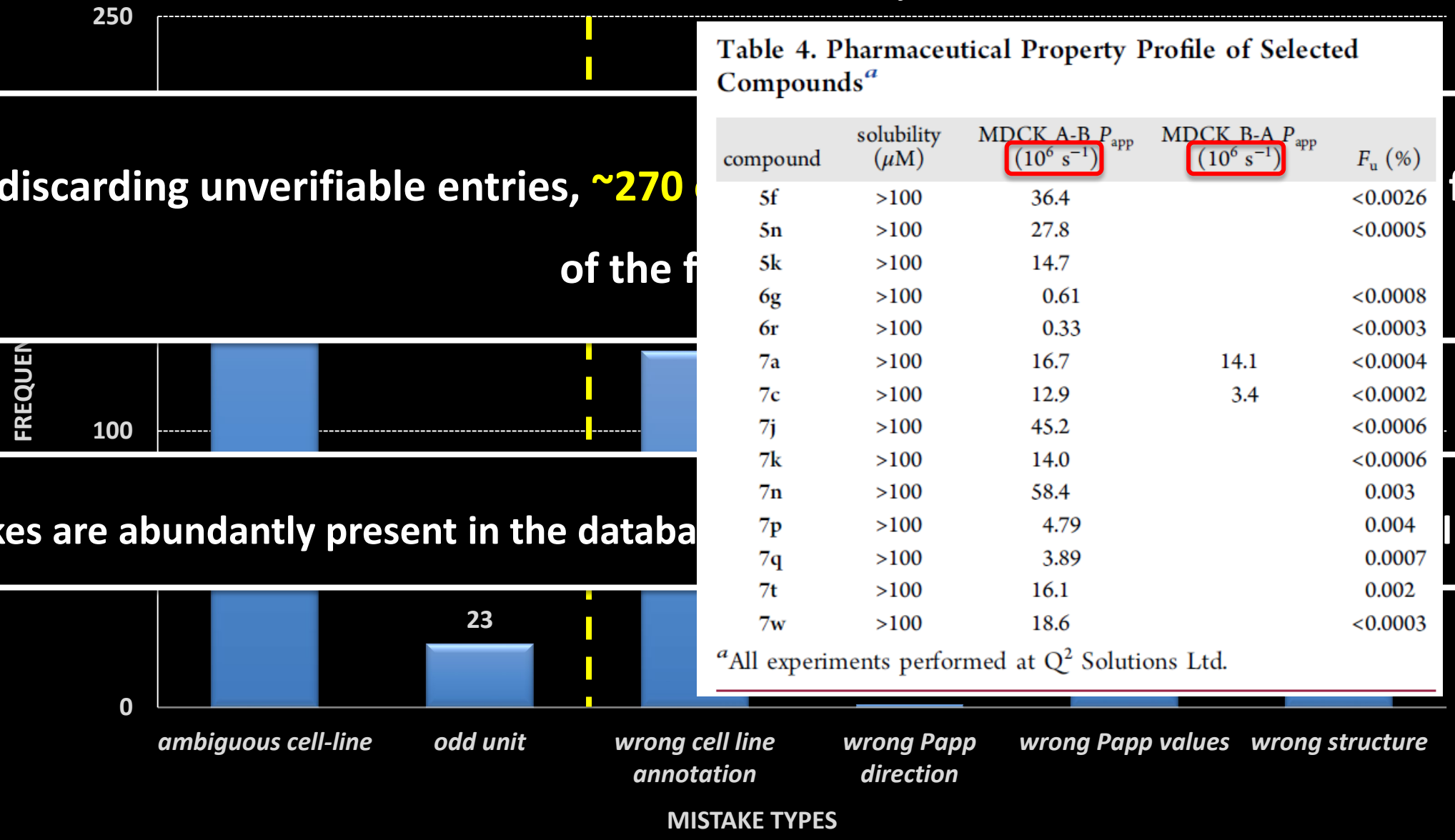


Table 4. Pharmaceutical Property Profile of Selected Compounds[a]

compound	solubility (μM)	MDCK A-B P_{app} (10^6 s^{-1})	MDCK B-A P_{app} (10^6 s^{-1})	F_u (%)
5f	>100	36.4		<0.0026
5n	>100	27.8		<0.0005
5k	>100	14.7		
6g	>100	0.61		<0.0008
6r	>100	0.33		<0.0003
7a	>100	16.7	14.1	<0.0004
7c	>100	12.9	3.4	<0.0002
7j	>100	45.2		<0.0006
7k	>100	14.0		<0.0006
7n	>100	58.4		0.003
7p	>100	4.79		0.004
7q	>100	3.89		0.0007
7t	>100	16.1		0.002
7w	>100	18.6		<0.0003

[a]All experiments performed at Q^2 Solutions Ltd.

Where's Waldo?

- ***How many "Waldos" are in the chemical dataset?***
  - ***It would be unrealistic to be 100% confident about dataset correctness, especially in very large datasets, but the goal is to find as many mistakes as we can using the tools and methods we have.***
- ***What types of "Waldos" are there?***

= Mistakes in the dataset

Mistakes such as

- *Biological Endpoints (Papp values)*
- ***Reported Units***
- ***Chemical Structures***
- ***Cell-line Information***
- ***Direction of permeability***

Inconsistency in Papp Unit Formats Reported

Physical properties for compounds **11–13**

Compd	$\log D_{7.4}$[a]	Mouse PPB[b] (% free)	Rat PPB[b] (% free)	Solubility[c] (μM)	MDCK permeability[d] Papp ($\times 10^{-6}$ cm s^{-1})	CYPS[e] IC_{50} (μM)	hERG[f] IC_{50} (μM)	Rat Heps Clint (μL/min/ 10^6 ells)

Source of error: Database

Type of error: Unit Formatting Inconsistency from Literature

Detection : Distribution Analysis

2			15	0.033
44			5	0.007
49	70	7	12	0.028
51	57	13	1	0.015
53	25	10	13	0.011
54	48	14	15	0.017

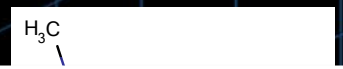
Corrected to 10^{-6} cm/s

IC_{50} μM (bioch)	0.04	0.06	0.07	0.09
IC_{50} μM (cell)	0.11	0.07	0.64	0.55
AlogP	2.3	3.4	2.4	2.1
MLM (mL/min/kg BW)	65, 109	76	93	77
MDCK EER	3.0	1.2	3.2	4.9
Papp A-B (cm/s x 10^6)	280	294	234	175

“Waldo”s in Papp Values

Source of error: Database
(unit conversion mistakes; common)

Source of error: Database (decimal mistakes; common)



Source of error: Database
Type of error: Human Mistake

Detection : Distribution Analysis, Duplicate Structure Assessment

ylcarbamoyl)-5-phenyl-pentyl]-4-pyridin-3-ylmethyl-
piperazine-2-carboxylic acid tert-butylamide
(Database) MDCK Papp = 0.87 nm/s
(Literature) MDCK Papp = 0.87 x 10^{-6} cm/s
DOI: 10.2174/138920007782109733

(Database) MDCK Papp = 0.000301 cm/s
(Literature) MDCK Papp = 30 x 10^{-6} cm/s

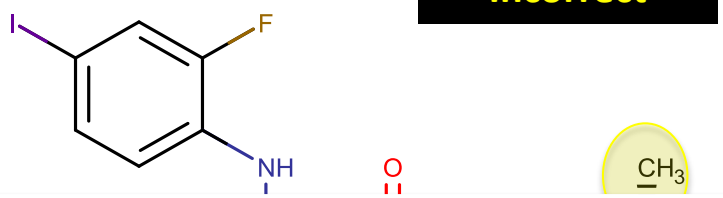
DOI: 10.1016/j.bmcl.2009.12.071

(Database) MDCK Papp = 0.000131 cm/s
(Literature) MDCK Papp = 13 x 10^{-6} cm/s

DOI: 10.1016/j.bmcl.2009.12.071

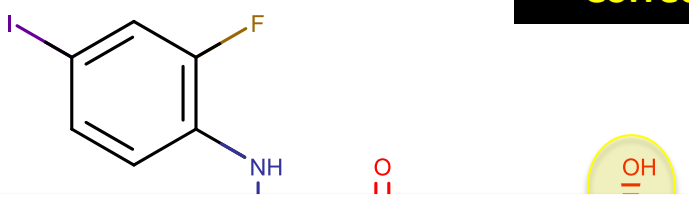
“Waldo”s in Structural Mistakes

Incorrect



Source of error: Database

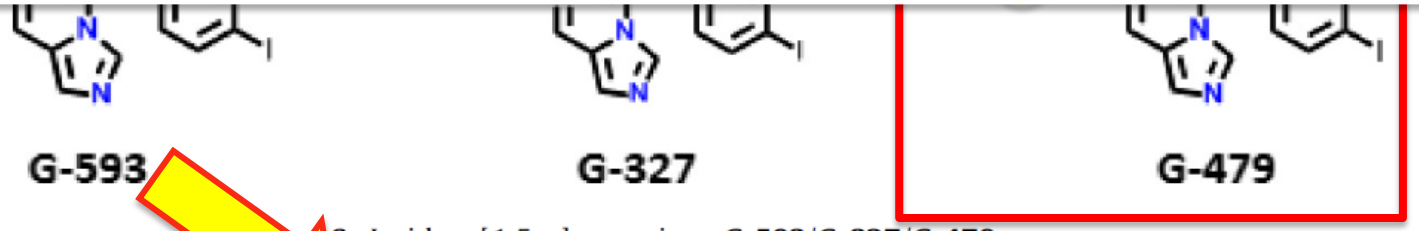
Correct



Source of error: Database

Type of error: Human Mistake

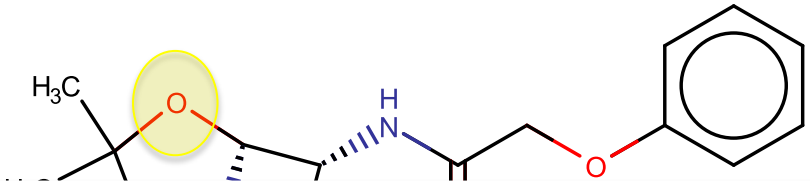
Detection : Duplicate Endpoint Assessment



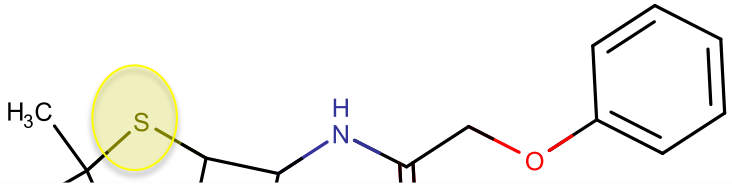
6. Imidazo[1,5-*a*] pyrazines **G-593/G-327/G-479**.

10.1016/j.bmcl.2014.08.008

“Waldo”s in Structural Mistakes



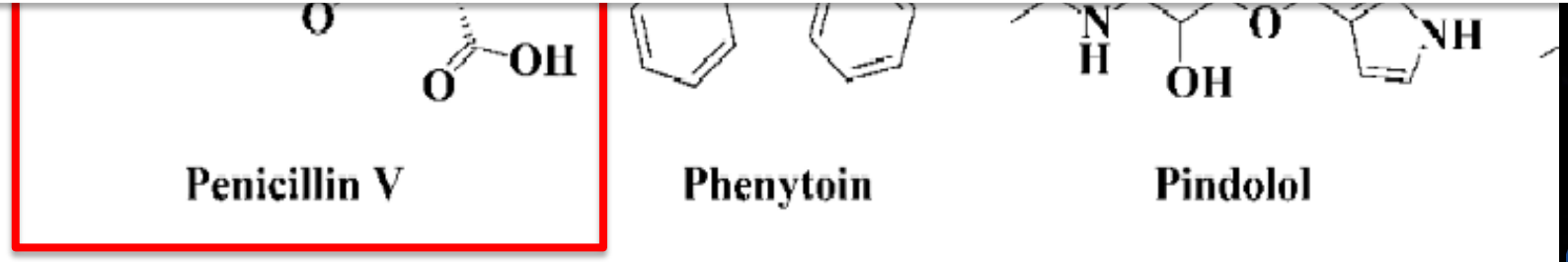
*Source of error: Literature



Source of error: Literature

Type of error: Human Mistake

Detection : Structural Verification Script



10.1111/j.1745-7254.2005.00166.x

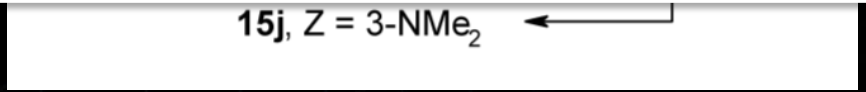
“Waldo”s in Structural Mistakes

https://pdfs.semanticscholar.org/0d4b/4786970a447a627c34c7362dd072a8a8d450.pdf

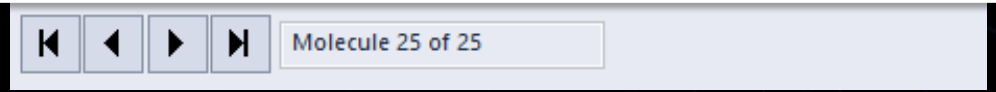
Source of error: Database

Type of error: Human Mistake

Detection : Cluster Analysis



Compound 15e
Papp (MDCK) = 0.78 ucm/s

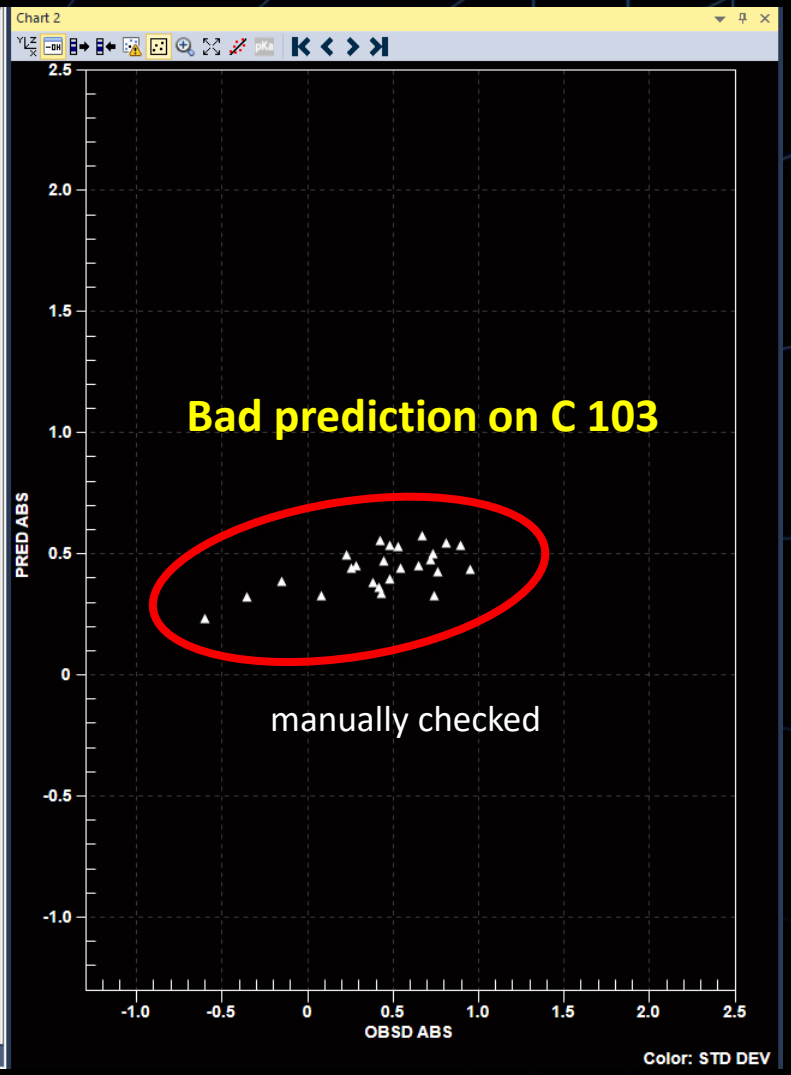


Compound 15e
Papp (MDCK) = 0.78 ucm/s

“Waldo”s in Structural Mistakes

	Clas...	Representative S...	Identifier	Class Size	Dist(med...	PRED ABS	OBSD ABS	STD DEV	activity_clif
1	C 1		Compou...	42		1.078	1.192	0.213	
2	C 7		CHEMBL...	39		1.043	1.106	0.197	
3	C 93		CHEMBL...	28		0.643	0.524	0.222	yes
4	C 40		Compou...	28		0.538	0.645	0.255	
5	C 103		CHEMBL...	25		0.438	0.432	0.172	yes
6	C 53		Compou...	25		1.019	0.996	0.174	
7	C 17		Compou...	25		0.256	0.150	0.272	
8	C 298		CHEMBL...	25		0.711	0.807	0.237	

Compounds | Classes | R Tables | Pairs | Keys



Conclusions

- **Several mistakes exist in chemical & biological databases and literature.**

  (Biological Endpoints, Reported Units, Chemical Structures, Cell-line Information, Direction of permeability, etc.)

- **These mistakes can be detected using a set of methodical, creative approaches based on the project's nature.**
  - Distribution analysis
  - Duplicate endpoint and structural analysis
  - Cluster analysis
  - Structure verifications
  - Manual inspection

- **A lot of effort must go into mistake detection and diagnosis part of the data curation process.**

Model-Informed Drug Development

MIDD+

2021 Virtual Conference

Q&A

Questions & Answers

phyophyo@simulations-plus.com

Learn More! www.simulations-plus.com

SimulationsPlus

Cognigen | DILIsym Services | Lixoft