MedChemica

CREATING A STEP CHANGE IN MEDICINAL CHEMISTRY

Practical Drug Discovery using Explainable AI

Dr Alexander G. Dossetter
Managing Director, MedChemica Ltd

Available on Slideshere - search for Dossetter
https://www.slideshare.net/AlDossetter/practical-drug-discovery-using-explainable-artificial-intelligence

Twitter @MedChemica
Twitter @covid_moonshot
Twitter #BucketListPapers
https://www.medchemica.com/bucket-list/

June2021

Agenda

- What is Explainable AI and What does it mean to MedChem?
- Designing AI systems to ***enable*** the chemist
- How do you get new ideas from a computer?
  - About Generative SMILE systems…..
  - Matched Molecular Pair analysis (MMPA)
    - Enumeration with MMPA
    - Drilling back to source data, to *explain* the origin of the Rules
- Explainable AI with ML Models
- Project Examples (6 projects including Covid Moonshot)
  - Please request extended version of the talk for your institute
- Future vision – A view on further AI supported chemistry

What is Explainable AI and What does it mean to MedChem?

June 2021

What is Artificial Intelligence (AI)?
What is Explainable AI?

- **The Turing test**, originally called the **Imitation game** by Alan Turing in 1950, is a test of a machine's ability to exhibit intelligent behaviour equivalent to, or indistinguishable from, that of a human.
- An Artificial Intelligence (AI) systems provides (or performs) *Actions*, either fully automated or with additional human experience and final decision making.
- AI ***does not*** mean Machine Learning (ML)
- Machine Learning (ML) is the collective term for algorithms that can analyse a dataset to create a 'model' that can perform predictions or generate options [they save humans from having to work out the 'Rules' governing a dataset and custom write a program]
- The 'models' (produced by ML and other techniques) are often called 'blackbox' – we, as humans, have no idea how they work…
- Explainable AI are systems (or models) where the human can "see" how they work and can link the Actions back to original data
  → ***For chemists it means we can see (sub)structures / measurements***

The 5 Levels of AI

Level	Label	Description
Level 5	Fully Automated	Never requires human intervention
Level 4	Automation	Runs itself unless it hits an 'extreme' situation
Level 3	Semi-Autonomous	Running and monitoring a system – e.g. auto trading on the stock exchange
Level 2	Close supervision	Routine stuff administered uses humans for difficult situations
Level 1	Simple Augmentation	Data entry, processing, identifying cluster of activity, profiling
Level 0	No automation	Human powered only

For Drug Discovery we can see that somewhere between Level 1 and Level 2 would be very useful given the current high volume of data and diversity

Designing AI systems to ***enable*** the chemist

A bit about MedChemica and how the computer can work with the chemists

June 2021

…9 Years of working with pharma companies

"Our median number of compounds per LO project is 3000 - this is unsustainable… [it should be] 300"

– Director of Chemistry (large pharma)

"Can we define the text book of medicinal chemistry?"

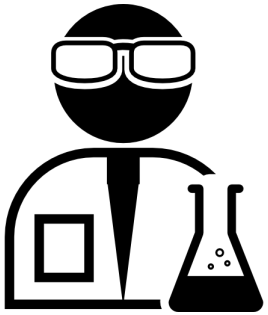
– Director of Comp Chem (large pharma)

"We are aiming at 300 compound per project – currently we are about 400, we will get better"

– ExScienta scientist at SCI 'What can BigData do for chemistry' – London Oct 2017

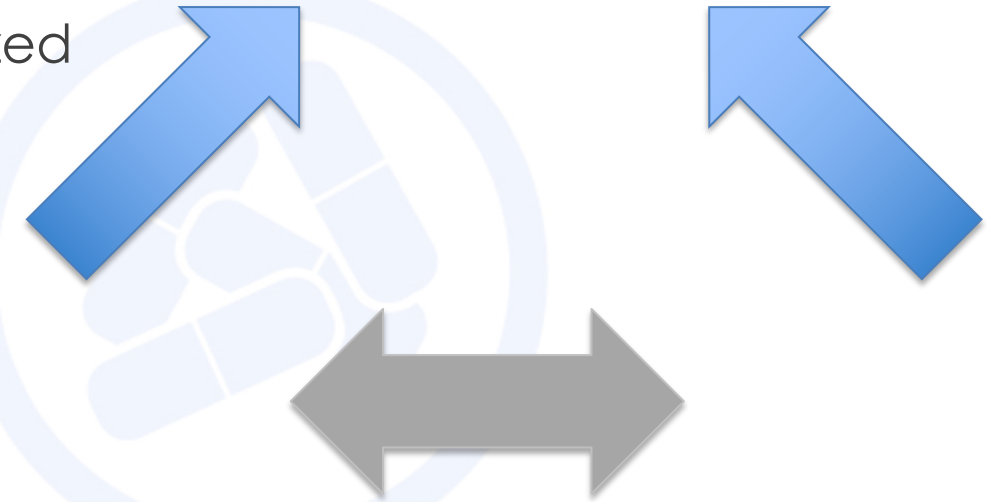
MedChemica: Using knowledge extraction techniques to build Augmented Intelligence systems to increase success and reduce the time and cost in Drug Discovery.

Augmenting the Medicinal Chemist

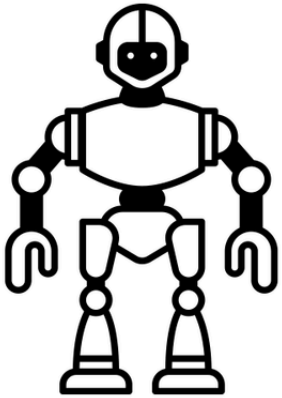


Sets goals
Makes Decisions

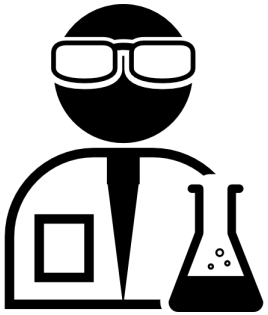
Data is organized and summarized



Prioritizes options



Augmenting the Medicinal Chemist

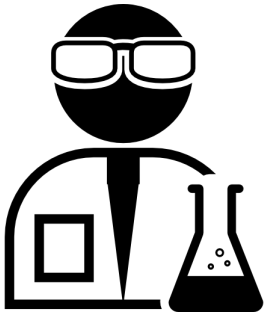


Sets goals
Makes Decisions

Data is organized and summarized

Data normalization
Pattern Recognition
Visualisation

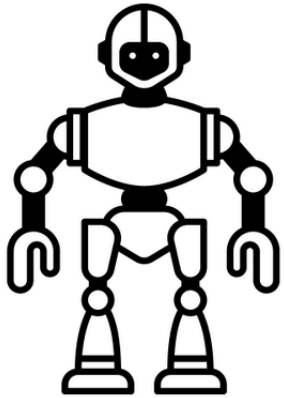
Augmenting the Medicinal Chemist



Sets goals
Makes Decisions

Makes proposals from information + goals
Automated improvement & Active Learning ….. *Situational awareness?*

Prioritizes options



What does AI mean to Medicinal Chemistry?

- So it is more helpful to consider Augmenting Intelligence in Medicinal Chemistry….

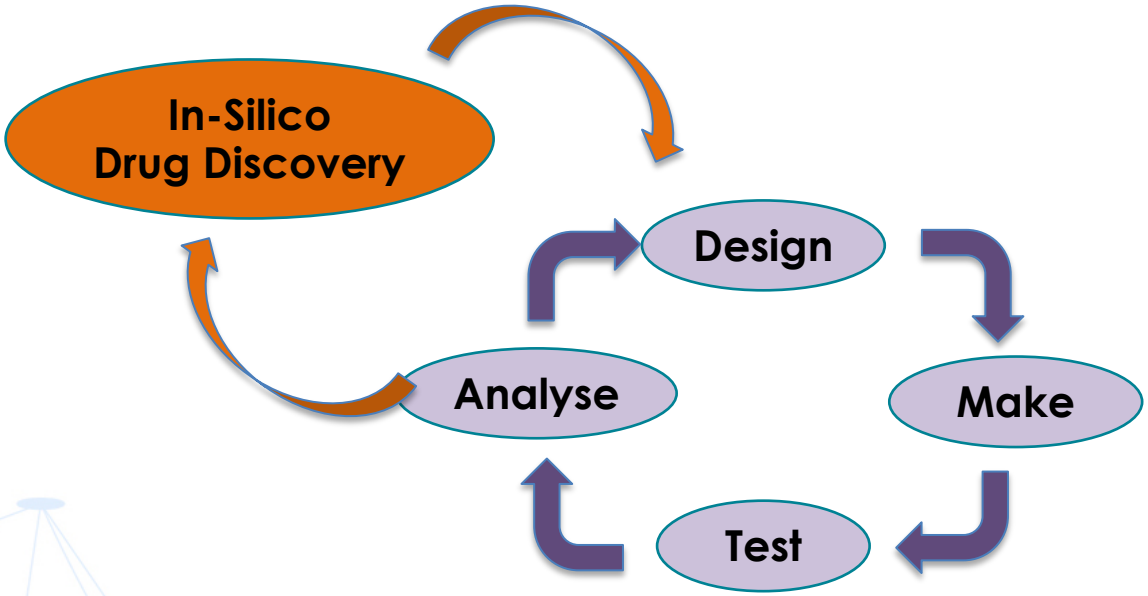
- The quality of an AI model depends on the number of times the machine can learn from success and failure.
  - Alpha-Go
    - fast to learn because a computer can play another computer
    - Clear success and failure
  - Drug Discovery 
    - Fully documented discovery projects (all cmpds/all measurements) in short supply
    - DTMA too long for many iterations?
    - Unclear success / failure in early research

"*Can we accelerate medicinal chemistry by augmenting the chemist with BigData and Artificial Intelligence?*" Griffen E.J. *et al* **Drug Disco. Today, 2018**, 23, 7, 1373-1384.

"Chemists: AI Is Here; Unite To Get the Benefits" Griffen, E.J.*; Dossetter, A.G.; Leach, A.G.; **J. Med. Chem. 2020**, 63, 16, 8695–8704 https://doi.org/10.1021/acs.jmedchem.0c00163

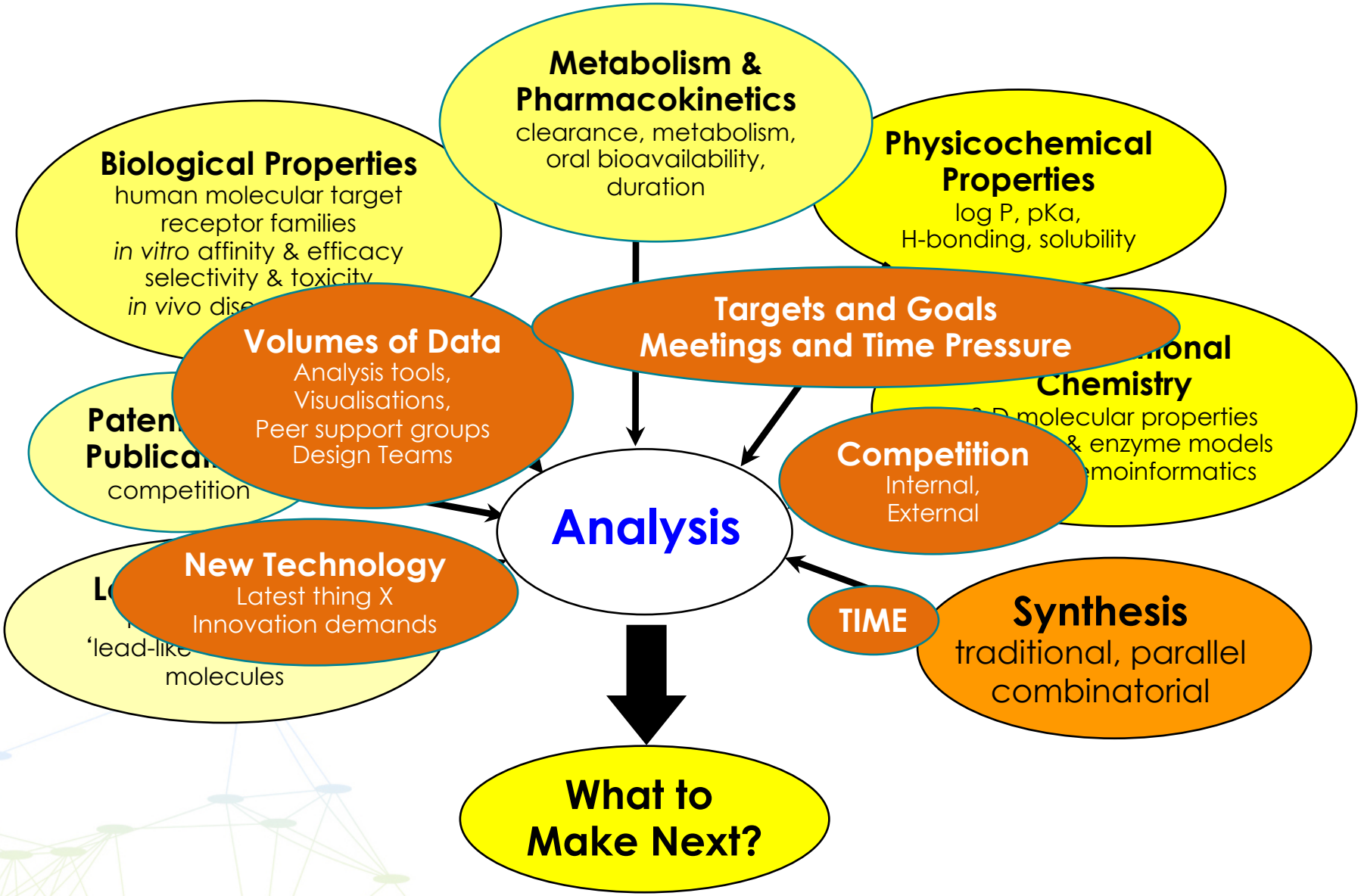
AI or In-Silico Drug Discovery

- In small molecule Drug Discovery Med Chem controls the productivity.
- Better Analysis and Design is the the only way to Making and Testing fewer compounds.
- In-Silico Drug Discovery covers both Analysis and Design
- AI techniques are being applied to improve In-Silico DD with unbiased analysis and recommendation systems (Level 1 and 2)



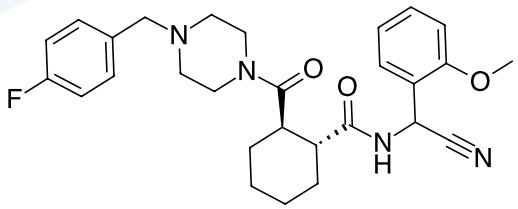
- **Omission** – New ML models have been built on synthetic chemistry routes to produce new route recommenders – not covered in this talk.

For the jobbing Med Chemist what does this mean?



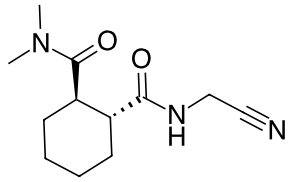
CD1 - Cathepsin K Inhibitors for OE – AZD4996

pIC50	8.70
LogD	3.6
HLM	127
Solubility	99µM
DTM	
Potent	
High Clearance / 0 F%	



ΔpIC$_{50}$	- 2.2
ΔLogD	- 4.1
ΔpSol	+3.0
ΔHLM	- 1.2

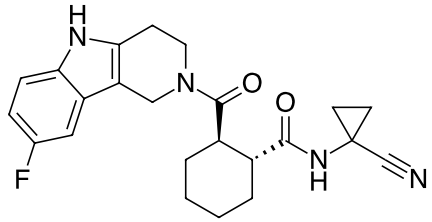
Molecular Simplification



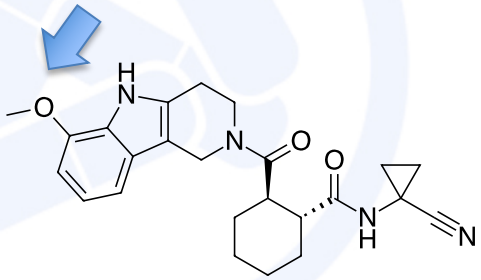
ΔpIC$_{50}$	+ 2.3
ΔLogD	+ 3.2
ΔpSol	- 1.0
ΔHLM	+ 1.0

43 Compounds made

Unusual structural change



ΔpIC$_{50}$	~ 0.0
ΔLogD	- 0.1
ΔpSol	+ 0.1
ΔHLM	- 0.4

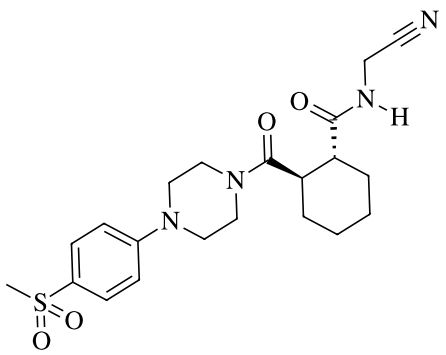


pIC50	9.1
LogD	2.8
HLM	<2.0
Solubility	>1000µM
DTM	0.05 mg/kg UID
High F% / stability maximised	

Dossetter, A.G. *et al Bioorg Med Chem Lett.* **2012**, *22*(17), 5563 - 5568.
Dossetter, A.G. *et al J Med Chem.* **2012**, *55*(14), 6363 - 6374.

CD2 - Cathepsin K Inhibitors for OE

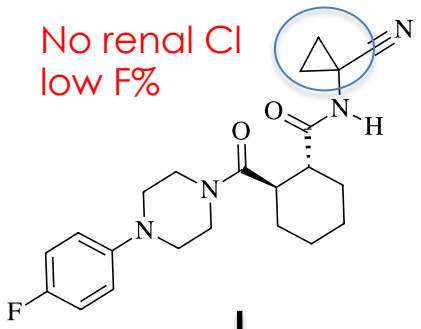
pIC50	7.95
LogD	0.67
HLM	<2.0
Solubility	280µM
DTM	~1.0 mg/kg UID
Potent	
Too polar / Renal Cl	



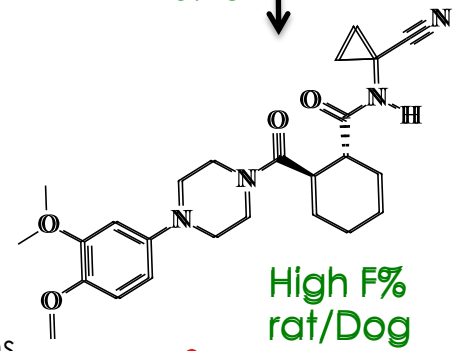
ΔpIC$_{50}$	- 0.1
ΔLogD	+1.4
ΔpSol	+1.2
ΔHLM	+ 0.25

Solubility

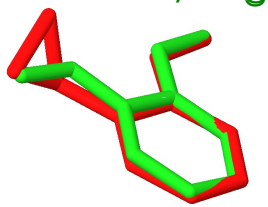
No renal Cl
low F%



ΔpIC$_{50}$	+0.1
ΔLogD	- 0.7
ΔpSol	~0.0
ΔHLM	- 0.25



High F%
rat/Dog

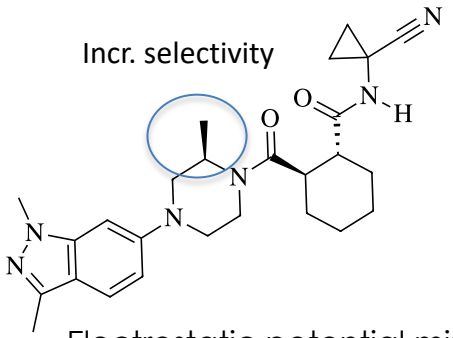


PDB - 97% of structures

Increase in LogP,
Properties improved

pIC50	8.2
LogD	2.8
HLM	<1.0
Solubility	>1400µM
DTM	0.01 mg/kg UID
High F% / stability maximised	

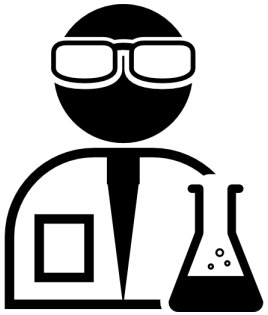
Incr. selectivity



Electrostatic potential minima between oxygens
Approx like N from 5-het, new compound can not form a quinoline

Crawford, J.J.; Dossetter, A.G *J Med Chem. 2012, 55, 8827.*
Dossetter, A. G. *Bioorg. Med. Chem.* **2010**, 4405

Augmenting the Medicinal Chemist

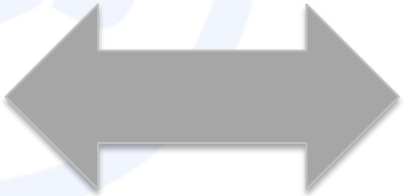


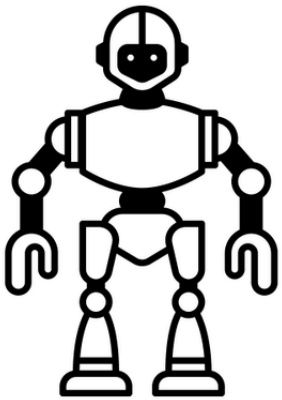
Sets goals
Makes Decisions

How do you get a computer to make sensible compound suggestions, aligned to a strategy, and where we can see where the idea has come from?

- *Generative SMILES models*
- *Matched Molecular Pair Models*

What to make next?



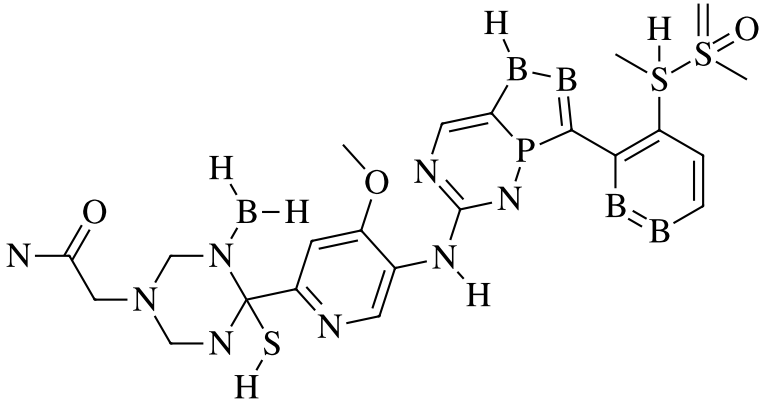


How do you get new ideas from a computer?

About Generative SMILE systems
Matched Molecular Pair analysis (MMPA)
Enumeration with MMPA
Drilling back to source data, to *explain* the origin of the Rules

About Generative SMILES models

- Compounds can be represented as SMILES notation e.g. benzene **`c1ccccc1`**
- Use ML (or DL) 'model' to how SMILES strings are constructed
- With the model we can then ask it produce 'novel' compounds
  - They can do this in their billions…..



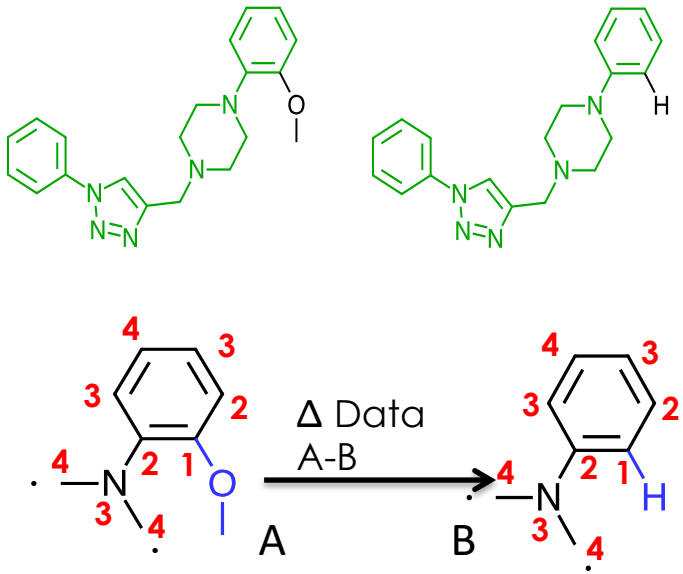
SMILES string

```
B1C(C(=BB2)P(C2=CN2)NC=2NC(=CN2)C(=CC=2C(S)(NCN2CC(=O)N)N(B)C2)OC)=C(C=CB=1)S(S(=C)(C)=O)C
```

De novo molecular design and generative models, Meyers, J.; Fabian, B.; Brown, N.; ***Drug Discovery Today*, June 2021**, https://doi.org/10.1016/j.drudis.2021.05.019

Fully Automated Matched Molecular Pair Analysis (MMPA) What is this form of Artificial Intelligence?

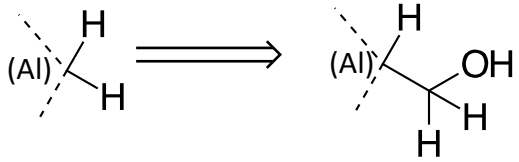
- Matched Molecular Pairs – Molecules that differ only by a particular, well-defined structural transformation
- Capture the change and environment – MMPs can be recorded as transformations from A→ B
- Statistical analysis to define "medicinal chemistry rules" Defined transformations with high probability of improving properties of molecules
- Store in a high performance database and provide an intuitive user interface

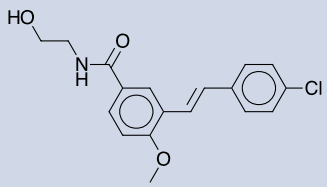
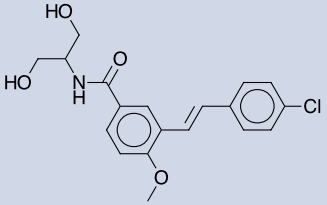
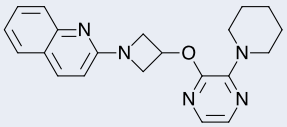
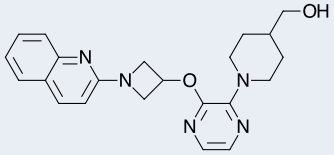
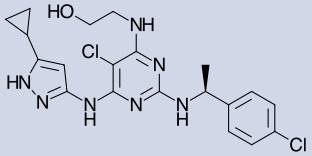
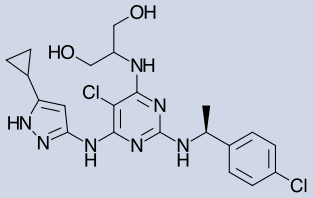


Griffen, E. *et al*. *J. Med. Chem.* **2011**, **54**(22), pp.7739 - 7750.

Leach *et al*. *J. Chem. Inf. Model.* **2017**, **57**, 2424 - 2436

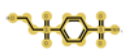
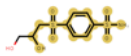
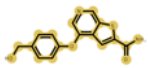
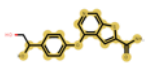
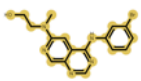
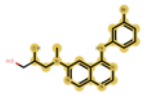
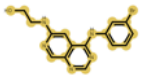
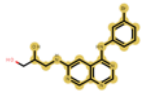
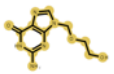
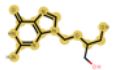
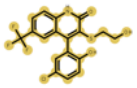
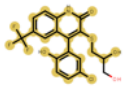
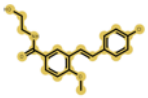
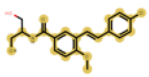
From SAR to MMPA.....

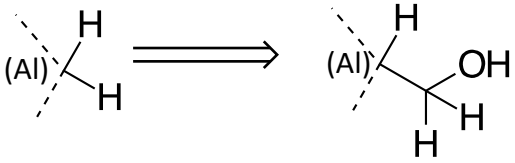


A	B	pSol A (µM)	pSol B (µM)	ΔpSol
 CHEMBL1949786	 CHEMBL1949790	- 4.3(48 µM)	- 3.2 (700µM)	1.1
 CHEMBL3356658	 CHEMBL218767	- 6.0 (1.0 µM)	- 3.7 (178 µM)	2.3
 CHEMBL456802	 CHEMBL456322	-5.7 (2.0 µM)	- 4.1 (82 µM)	1.6
				3 pairs +ve Sol Median 1.6

MCPairs Rule finder required 6 matched pairs for 95% confidence

The Matched Pairs leading to Rule.....

compound name A	compound name B	Depiction A	Depiction B	Aqueous Solubility at pH 7.4 solubility [CHEMBL2362975] unit	Aqueous Solubility at pH 7.4 solubility [CHEMBL2362975] qualifier A	Aqueous Solubility at pH 7.4 solubility [CHEMBL2362975] measurement A	Aqueous Solubility at pH 7.4 solubility [CHEMBL2362975] qualifier B	Aqueous Solubility at pH 7.4 solubility [CHEMBL2362975] measurement B	Aqueous Solubility at pH 7.4 solubility [CHEMBL2362975] measurement delta
CHEMBL104459	CHEMBL316800			log10(M)	=	-1.6042	=	-1.2971	0.3071
CHEMBL118022	CHEMBL115462			log10(M)	=	-4.63875	=	-4.3343	0.30445
CHEMBL161956	CHEMBL165547			log10(M)	=	-9.9586	=	-9.7447	0.2139
CHEMBL165864	CHEMBL166093			log10(M)	=	-10.699	=	-9.6778	1.0212
CHEMBL184	CHEMBL182			log10(M)	=	-5.2353	=	-4.7115	0.5238
CHEMBL184521	CHEMBL439660			log10(M)	=	-5.0168	=	-3.3704	1.6464
CHEMBL1949786	CHEMBL1949790			log10(M)	=	-4.31675	=	-3.1588	1.15795



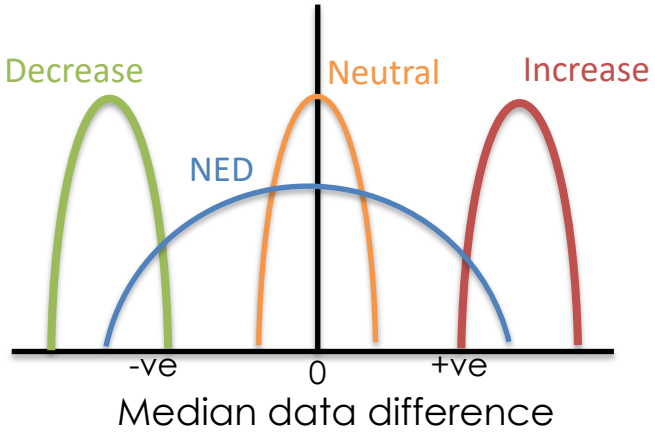
Actual Rule from MCPairs
Endpoint:
Aqueous Solubility at pH 7.4 [CHEMBL2362975]

n-qual	69
n-qual-up	47
n-qual-down	21
median ΔpSol	0.26
std dev +/-	0.636

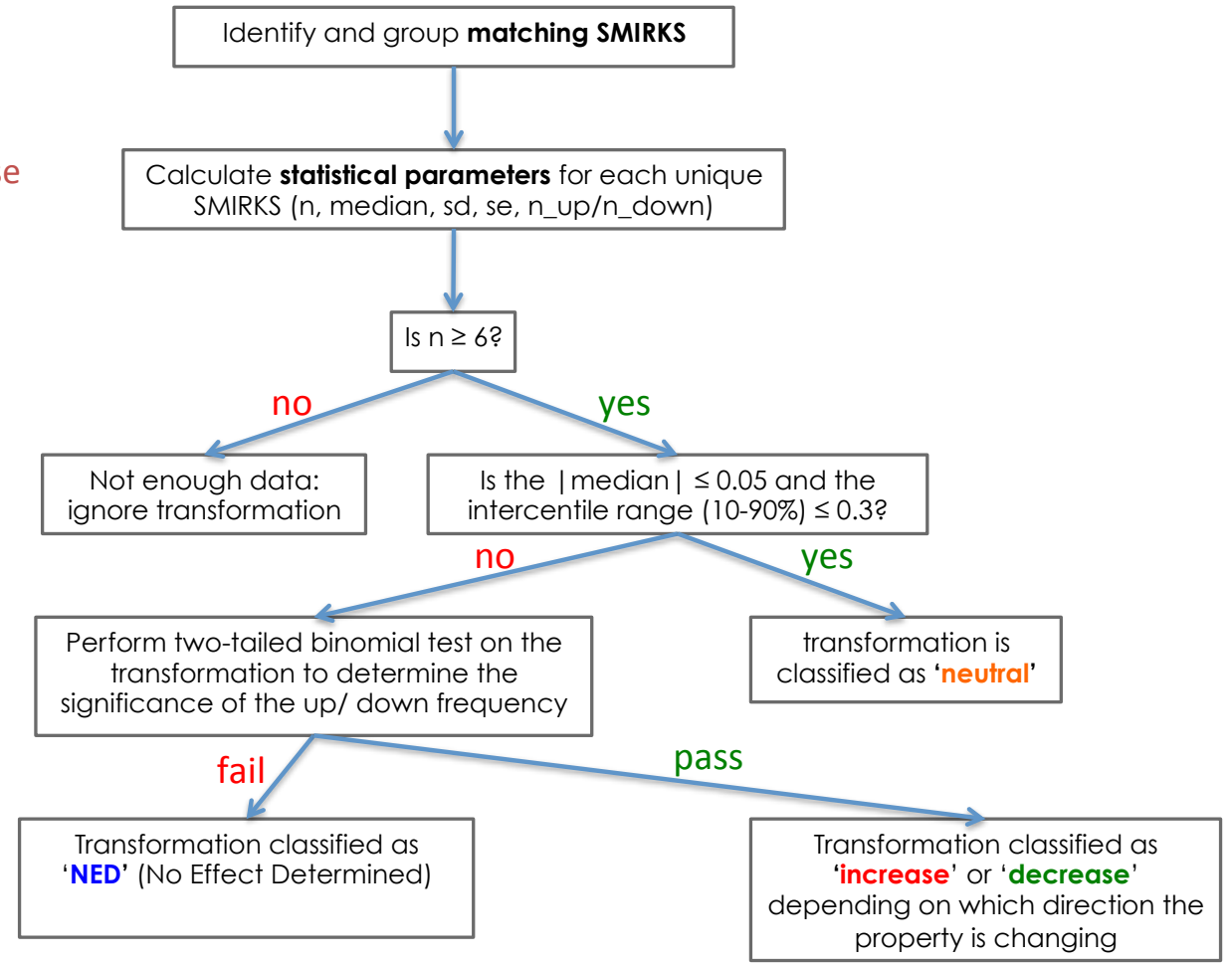
Explainable
- Drill back to real world examples and measured data

Actionable
- Clear decision to make the compound

Rule selection



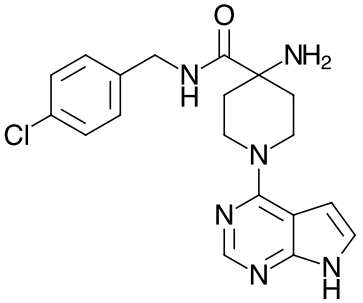
- No assumption of normal distribution
- Manages 'censored' = qualified / out-of-range data



Leach et al. *J. Chem. Inf. Model.* **2017**, **57**, 2424 - 2436

Capivasertib (AZD5363) – AKT inhibitors

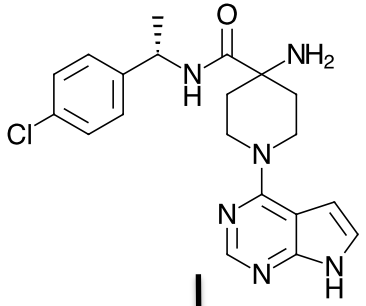
AKT pIC_{50} 7.89 (13nM)
LogD 2.9
Sol (pSol) -5.3 (5 – 150µM)
hERG pIC_{50} 5.2 (5.2 µM)
Potent enough
hERG and improved solubility



CHEMBL598194

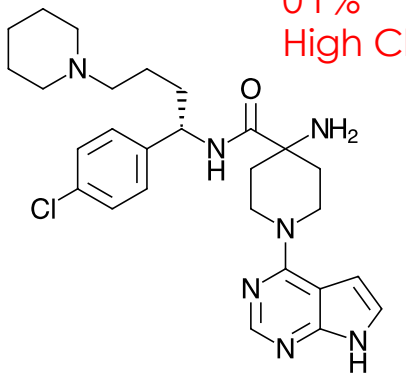
ΔpIC_{50}	+ 0.2
ΔLogD	- 0.2 (?)
ΔpSol	+ 0.5
ΔhERG	~ 0.2

CHEMBL2325742

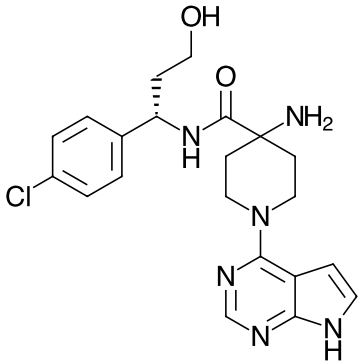


ΔpIC_{50}	~ 0.0
ΔLogD	- 0.3
ΔpSol	+ 1.4
ΔhERG	- 0.5

0 F%
High Cl



CHEMBL2325729

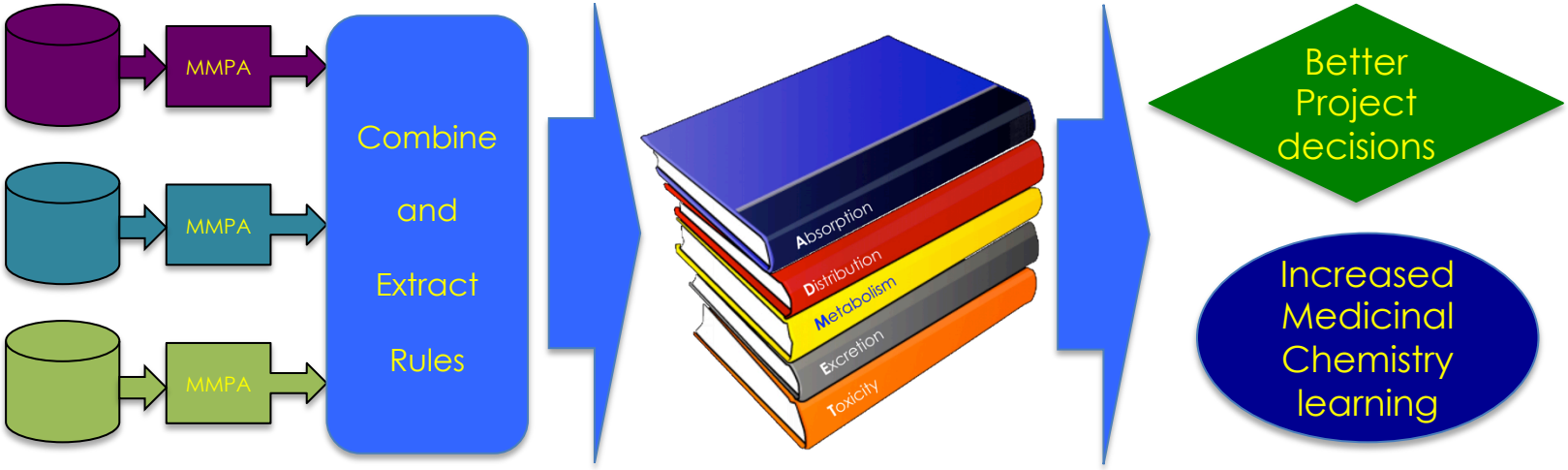


CHEMBL2325741

AKT pIC_{50} 8.4 (3nM)
LogD 2.5
Sol (pSol) -3.1 (780µM)
hERG pIC_{50} <4.0 (>100 µM)
Potent
No hERG and improved solubility

Kettle, J.G. *et al* ; J. Med. Chem. **2013**, **56**, 5, 2059–2073. http://dx.doi.org/10.1021/jm301762v

MMPA Enables knowledge sharing



Multiple Pharma ADMET data

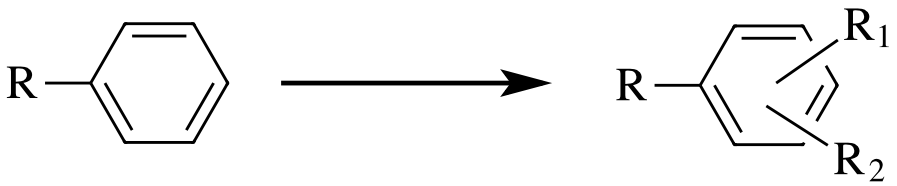
>437000 rules

Kramer, Robb, Ting, Zheng, Griffen, *et al*. *J. Med. Chem.* **2018, 61(8), 3277-3292**
http://pubs.acs.org/doi/10.1021/acs.jmedchem.7b00935

Our MMPA technology enabled knowledge sharing between multiple organisations (AstraZeneca, Hoffman La Roche and Genentech)

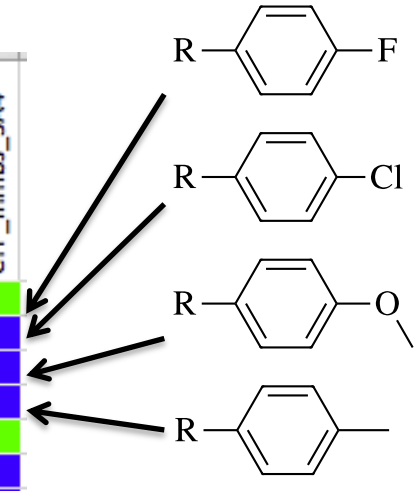
Most common phenyl substitutions

MedChemica
CREATING A STEP CHANGE IN MEDICINAL CHEMISTRY



* - Only the highest specificity data was used

Product_Group	no_of_MPs	LogD_TM	SOLUBILITY	PPB_HUM	PPB_RAT	Heps_HUM	Heps_RAT	LM_HUM	LM_RAT	hERG_inhib_HUM	CYP_inhibs_1A2	CYP_inhibs_2C19	CYP_inhibs_2C9	CYP_inhibs_2D6	CYP_inhibs_3A4
c1cc(cc[c]1)F	5894														
c1cc(cc[c]1)Cl	4431														
COc1cc[c]cc1	3595														
Cc1cc[c]cc1	2525														
c1cc(cc[c]1)C#N	1921														
c1cc(cc[c]1)O	732														
c1cc(cc[c]1)C(F)(F)F	684														
c1cc(cc[c]1)Br	512														
mono_sub_phenyl	472														
mono_sub_phenyl	392														
mono_sub_phenyl	383														
mono_sub_phenyl	278														
mono_sub_phenyl	237														
mono_sub_phenyl	223														
mono_sub_phenyl	222														
mono_sub_phenyl	208														
mono_sub_phenyl	159														
mono_sub_phenyl	153														
mono_sub_phenyl	129														
mono_sub_phenyl	106														
mono_sub_phenyl	103														



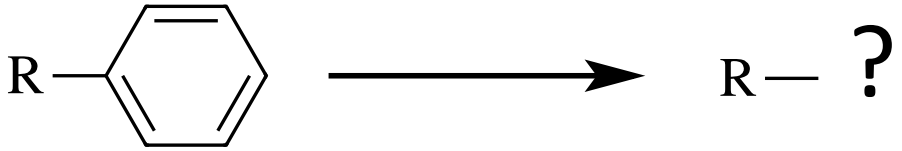
Mean >zero
Effect is good

Mean <zero
Effect is Detrimental

Mean ~ zero
Effect is neutral

No Effect Determined

Are there great changes from Phenyl?



"I want rules that fix a couple of problems but leave lots of things alone" (MPO)

Product_Group	LogD_TM	SOLUBILITY	PPB_HUM	PPB_RAT	Heps_RAT	Heps_HUM	LM_CYNO	LM_DOG	LM_HUM	LM_RAT	hERG_inhib_HUM	CYP_inhibs_1A2	CYP_inhibs_2C19	CYP_inhibs_2C9	CYP_inhibs_2D6	CYP_inhibs_3A4
heterocycle-1																
heterocycle-2																
mono-sub-phenyl																
mono-sub-phenyl																
heterocycle-3																
mono-sub-phenyl																
heterocycle-4																
heterocycle-5																
heterocycle-6																
heterocycle-7																
mono-sub-phenyl																

Mean >zero
Effect is good

Mean <zero
Effect is
Detrimental

Mean ~ zero
Effect is neutral

No Effect
Determined

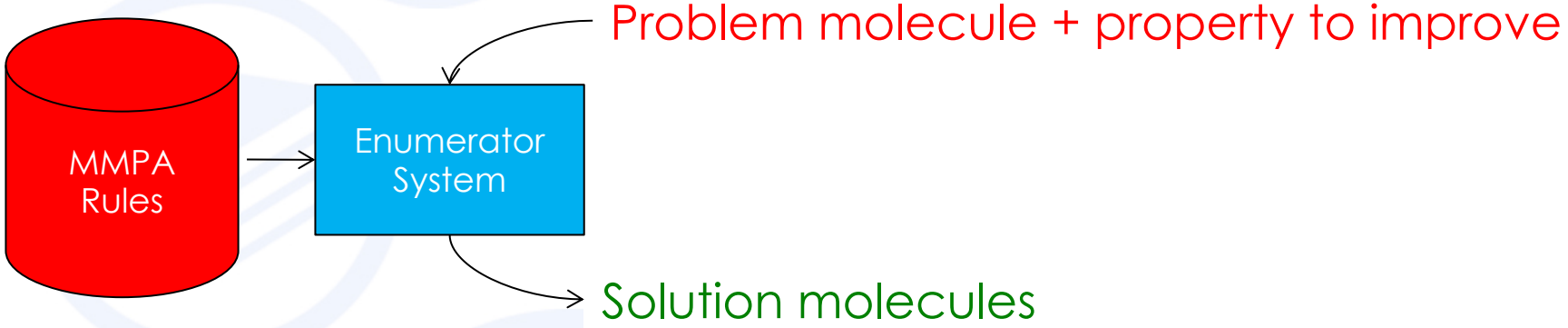
There are 25 thousand reasonable heterocycles for Med Chem

Tu, M. et al. J. Chem. Inf. Model. (2012), 52, 1114–1123

* - Only the highest specificity data was used

MCPAIRS ENTERPRISE

RuleDesign®



RuleDesign® (enumerate "Compounds From Rules")

- User enters in a sub-optimal molecule with a property they wish to improve – e.g. solubility, metabolism, hERG….
- System suggests new molecules considering the context of the changes **(Level 1 / 2)**

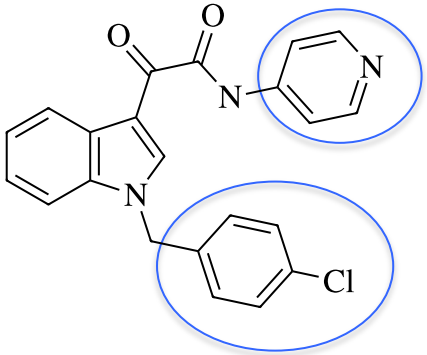
"..it's like asking 150 of your peers for ideas in just a few seconds"
- Principal Scientist (large pharma)

A Turing test for molecular generators Darren Green D.; *et al*
J. Med. Chem. 2020, 2020, 63, 20, 11964–11971

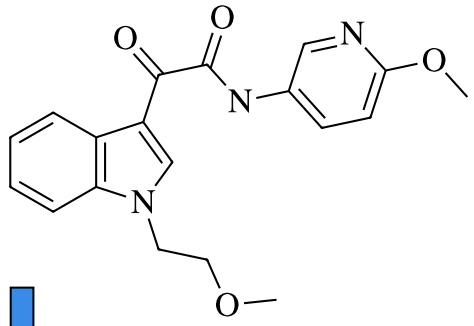
Indole-3-glyoxylamide Based Series of Tubulin Polymerization Inhibitors

- Increase potency, solubility and reduce metabolism
- Enable in-vivo xenograft studies

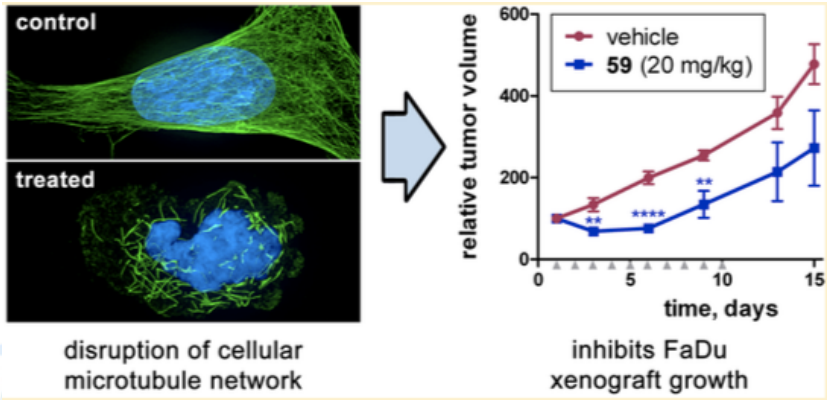
Indibulin D-24851
LC50 **0.032**
XlogP 3.35
~ potent
In-vivo activity
poor solubility (~ 1uM)



MMPA solubility & QSAR calcs

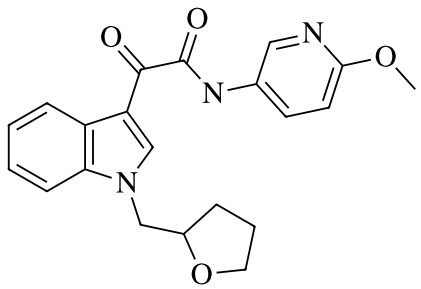


LC50 **0.027**
XlogP **2.02**

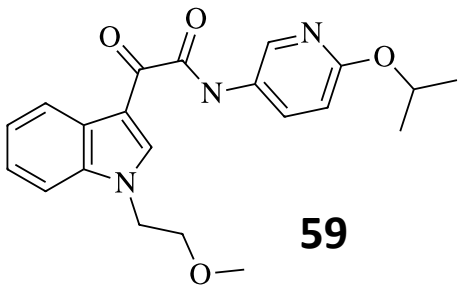


disruption of cellular microtubule network

inhibits FaDu xenograft growth



LC50 **0.031**
XlogP **2.57**
solubility (~10-80uM)



59

LC50 **0.055**
XlogP **2.91**
solubility (~10-80uM)

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

AI design → synthesis → In-vivo → publication 1 year

Journal of Medicinal Chemistry

Article

pubs.acs.org/jmc

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.

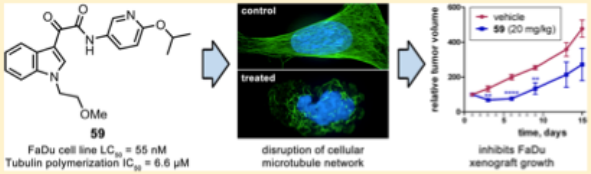
[§]Academic Unit of Clinical Oncology and Sheffield Experimental Medicine Centre, Weston Park Hospital, Whitham Road, Sheffield S10 2SJ, U.K.

[∥]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.

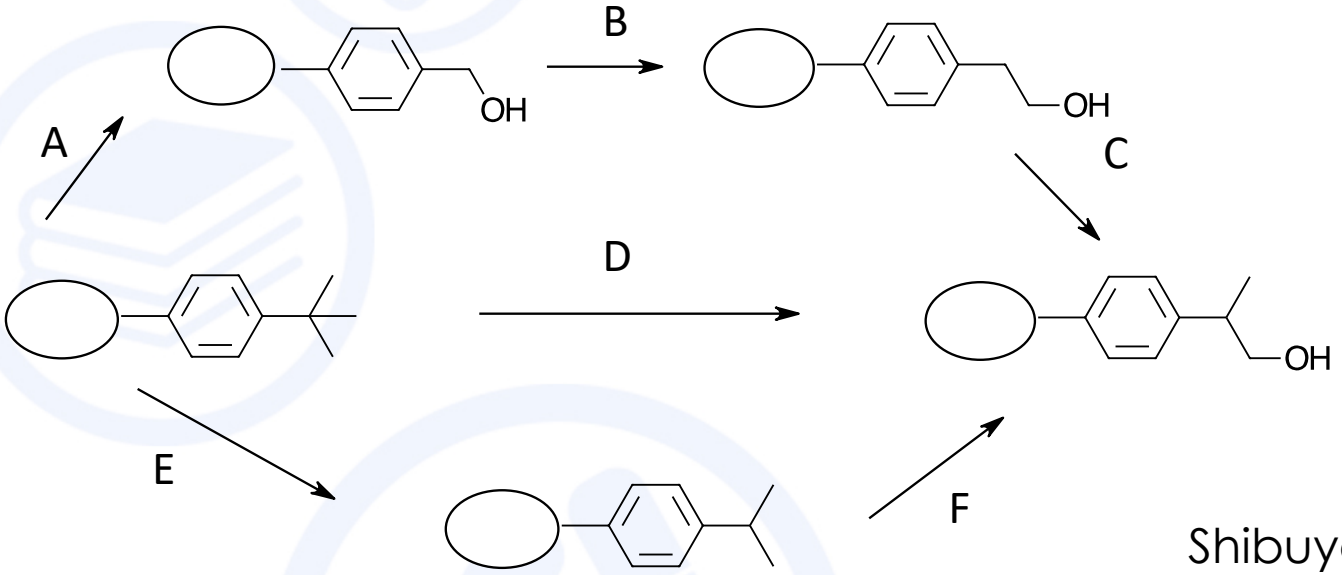
Supporting Information



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.; *J. Med. Chem.*, **2015**, 58 (23), 9309 – 9333

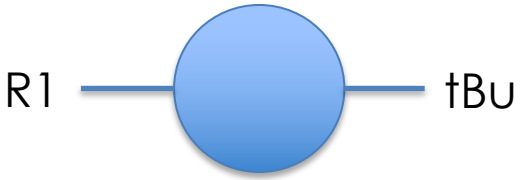
"Multi-Step" transformations

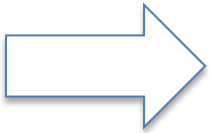


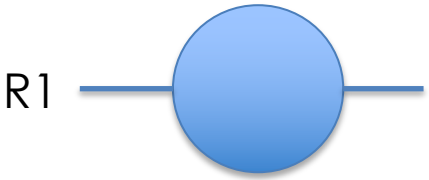
Would you go steps via A -> B -> C
How would you go know to go E -> F
Or go straight there via D
- if the data said it was good?

Shibuya Crossing Tokyo

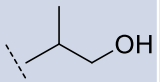
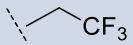
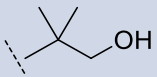
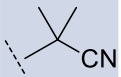
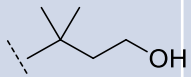
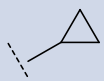
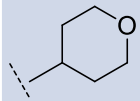
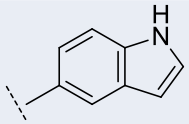
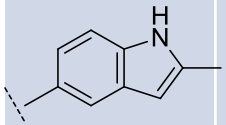
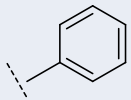
Example - ^{t}Bu metabolism issue







Roger Butlin
Rebecca Newton
Allan Jordan

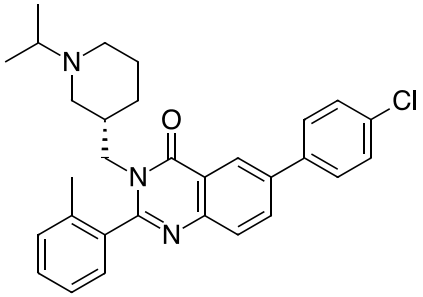
	Benchmark compound	Predicted to offer most improvement in microsomal stability (in at least 1 species / assay)									
R2 / R1	tBu				Me				Et	iPr	
	99 392	16 64	78 410	53 550	99 288	78 515	41 35	98 327		92 372	24 247
	35 128				24 62				60 395		
	39 445	3 21			20 27			57 89		54 89	

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

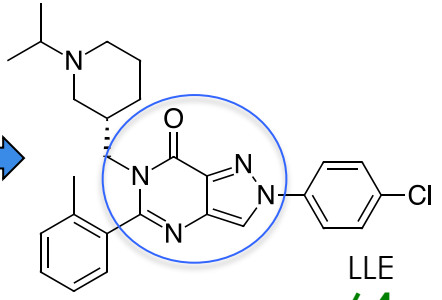
Project 4 – Ghrelin Inverse agonists – CNS target

- Novel more efficient core required, improve hERG for CD
- CNS penetration, good potency and deliver tool for in vivo testing

pIC50	9.9
logD	5.0
hERG pIC5	5.0
LLE	**4.9**
very potent	
very lipophilic	



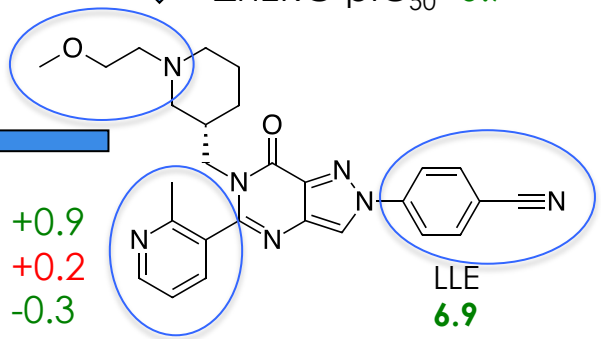
MMPA Cores



LLE **6.4**

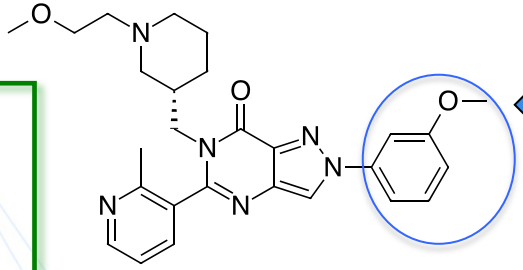
ΔpIC_{50}	-0.4
$\Delta logD$	-1.8
$\Delta hERG\ pIC_{50}$	+0.4

ΔpIC_{50}	-2.2
$\Delta logD$	-2.2
$\Delta hERG\ pIC_{50}$	-0.7



LLE **6.9**

ΔpIC_{50}	+0.9
$\Delta logD$	+0.2
$\Delta hERG\ pIC_{50}$	-0.3



100 compounds made

pIC50	8.2
logD	1.3
hERG pIC50	4.4
LLE	**6.9**

McCoull, W.M.; Dossetter A.G.; *et al,* Med. Chem. Commun., (**2013**), **4**, 456

LLE = lipophilic ligand efficiency:
LLE=pIC50-logD

What makes MMPA Explainable?

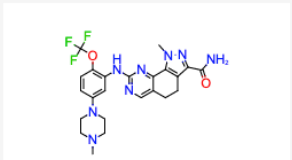
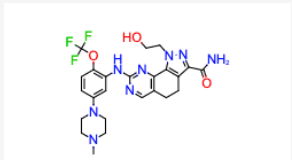
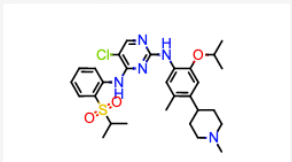
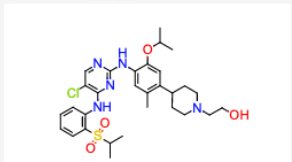
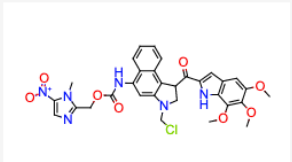
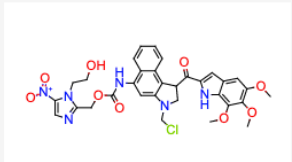
MedChemica
CREATING A STEP CHANGE IN MEDICINAL CHEMISTRY

Rule to Pairs

Modules | My Account

Transformation Information

Hide Structures

A	B	...	Aqueous Solubility at pH 7.4 solubility [CHEMBL2362975] uM				
			Qual A	A	Qual B	B	Fold Change
CHEMBL1272140 	CHEMBL1738758 		=	39.264	=	84.850	~ 2.2
CHEMBL2403849 	CHEMBL2403836 		=				
CHEMBL356820 	CHEMBL153380 		=				
CHEMBL3357064	CHEMBL3357070						

In just a few clicks we arrive back at the original compounds and measured data to see where the Rule has come from. Fully explained because we can see the structures!

Explaining AI with ML Models

It is possible to produce a potency prediction with ML models
AND
show the chemist how that prediction is derived

What does Explainable AI mean?

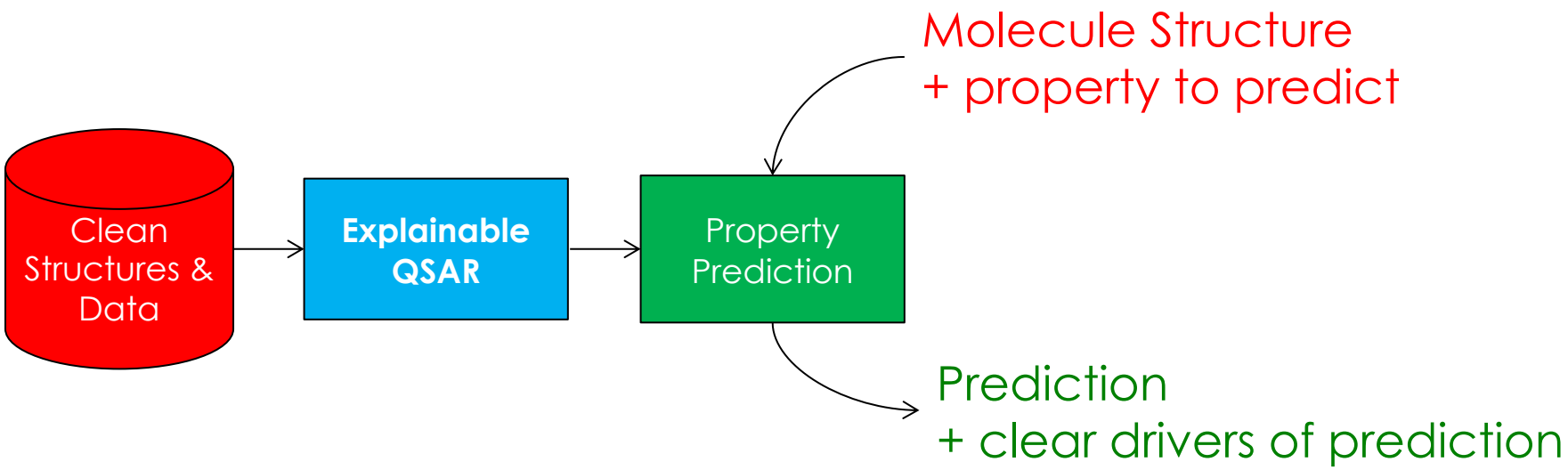
About compound descriptors and Methods…

Descriptors / Method	substructures	Physical chemistry descriptors(Hansch, Taft, Fujita, Abraham)	Atomic, pair, triplet descriptors	Indices
(M)LR				
PLS				
Trees / Forests				
SVM				
Bayesian NN				
Deep Learning				

INTERPRETABILITY

Increase clarity?

Property Prediction and Explainable AI Models



Automated Explainable QSAR

Chemists get predictions with the substructures highlighted that are driving prediction and the molecules used to support that part of the model – transparent / explainable AI.

Explainable ML Models

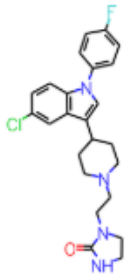
MedChemica
CREATING A STEP CHANGE IN MEDICINAL CHEMISTRY

Muscarinic acetylcholine receptor M1 Homo sapiens pIC50 [CHEMBL216]

c1cc(ccc1n2cc(c3c2ccc(c3)Cl)C4CCN(CC4)CCN5CCNC5=O)F

Explainable
- Highlighted features show the chemist the contribution to the prediction

Actionable
- Which parts should be optimized to achieve the Goal

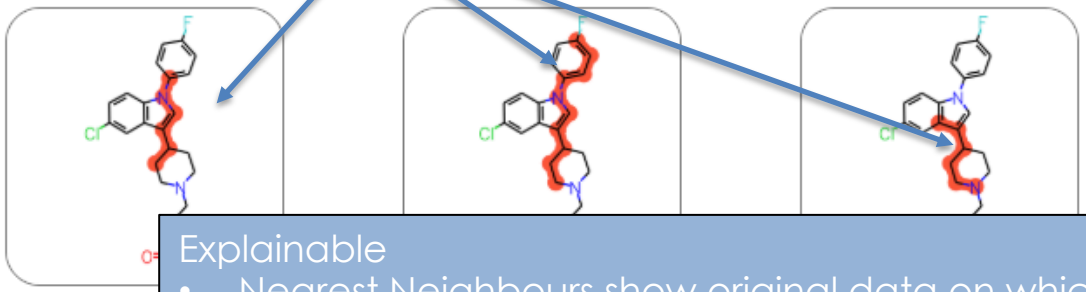


3 Pharmacophore(s)

Active

In domain

uM: 0.096
Range: 0.321 - 0.028



LH_4ato
median with [witho
n with / total:

Explainable
- Nearest Neighbours show original data on which model is built

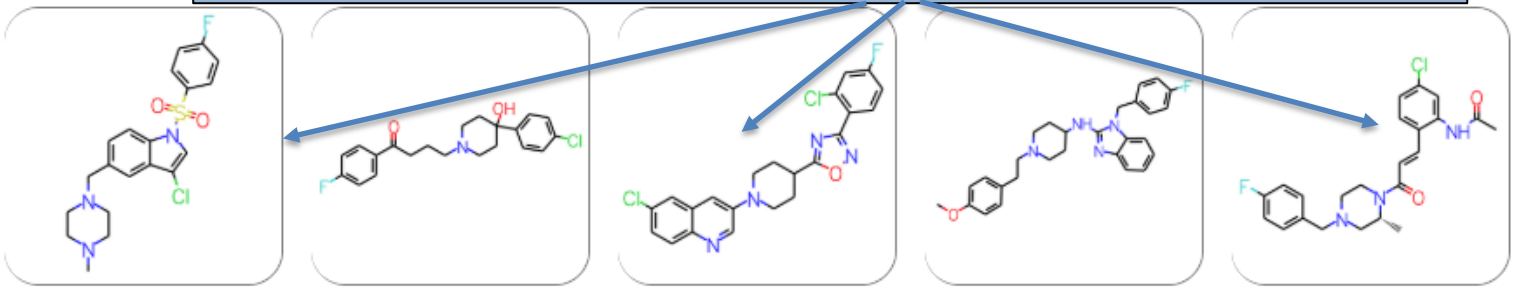
Actionable
- What weight do I put on this results? How likely is it? Do we test?

10 Nearest Neighbour(s)

Active

Out of domain

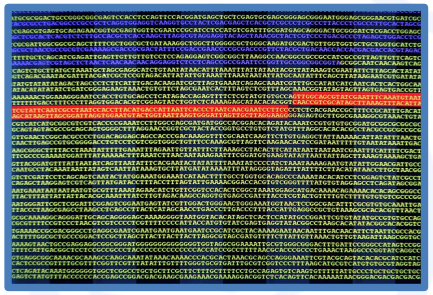
uM: 2.555
Range: 3.584 - 1.821



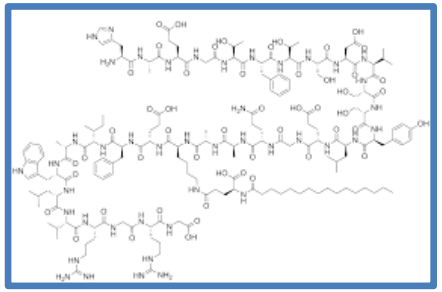
CHEMBL1785074	R-1625	CHEMBL468176	GNF-Pf-2461	CHEMBL198949
uM: 10.000	uM: 5.500	uM: 50.003	uM: 1.066	uM: 1.000
Tan. distance: 0.696	Tan. distance: 0.744	Tan. distance: 0.764	Tan. distance: 0.767	Tan. distance: 0.769

Future vision – A view on further AI supported chemistry

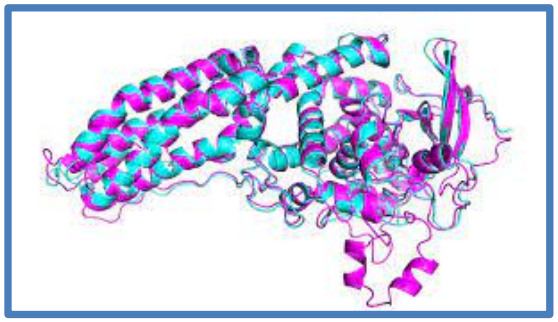
Future of Augmented Intelligence in Drug Discovery?



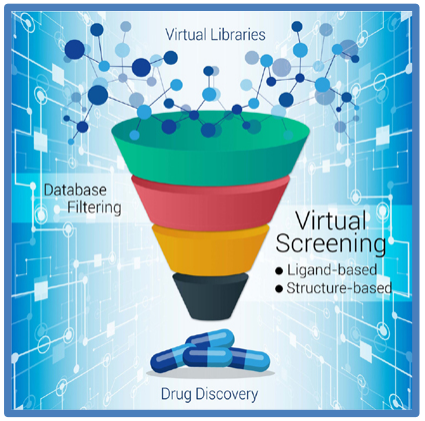
Genome Sequence



Peptide sequence

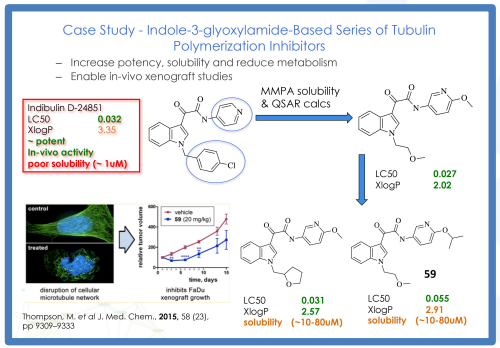


3D peptide folding
Identify binding pockets



Virtual Screen – 2000 hits

Biochemical Screening



Lead Optimisation
100 compounds

THE COVID MOONSHOT

MedChemica
CREATING A STEP CHANGE IN MEDICINAL CHEMISTRY

Goal: new potent antiviral: therapeutic & prophylactic

- *simple synthesis*
- *orally potent, 3x daily for 5 days*
- *very safe*
- *simple tablet*

Strategy: work fully open to enable rapid global availability

- *data and compound designs immediately public*
- *no IP filed*
- *generic drug "straight from pipeline"*
- *(unprecedented – no template available)*

June

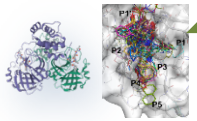
Feb 14th — *March 18th* — *Sep '20* — *Sept '21* — *Mar '22*

XChem → **fragment-to-lead** → **lead optimization** → **pre-IND** → **trials**

(planning) *(planning)*

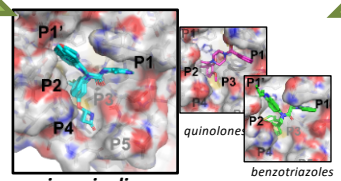
$350k, 30 groups *$800k, 40 groups*

Suite of optimized leads ***Regulatory phase***



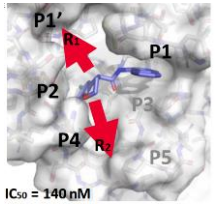
MPro cloned
Structure solved
78 fragment structures solved

Data release



isoquinolines

Lead series



Optimized leads

IND

7 months: 3 lead series
200nM enzyme inhibition
cellular antiviral activity
(philanthropic funding)

Current status: *potent but simple (<25nM)*
antiviral inhibition (113nM)
inhibits Variants of Concern
highly encouraging safety signals

Optimizing: rodent *oral exposure (blood levels)*

Establishing: strategy to pharmacy
pre-IND delivery
delivery partners
healthcare system
funding sources
ways to accelerate

https://www.diamond.ac.uk/covid-19/for-scientists/Main-protease-structure-and-XChem.html
https://www.nature.com/articles/d41586-021-01571-1

Conclusions

- What is AI, Explainable AI and What does it mean to MedChem?
  - Think Augmented Intelligence (Level 1 and 2)
  - Viewing (sub)structures and measurements *explains* the computer's thinking
- Designing AI systems to ***enable*** the chemist
  - Sensible compound suggestions with simple interfaces
- How do you get new ideas from a computer?
  - Matched Molecular Pair analysis (MMPA)
    - Suggests compounds with decisions made by the chemists (Level 2)
    - Drilling back to source data, to *explain* the origin of the Rules (Level 1)
    - Permutative MMPA to ensure all the gaps are filled (Level 2)
    - ML models with substructural features – where do we need to focus (Level 1)
- Project Examples (6 projects including Covid Moonshot)
- Future vision – A view on further AI supported chemistry
  - Invest in Data there is still lots of chemical space we do not understand
  - Compute methods are improving all the time…

Thank you

- Dr Alexander G. Dossetter
- ManagingDirector, MedChemica Ltd
- al.dossetter@medchemic.com

- **Available on Slideshere - search for Dossetter**

- **Twitter @MedChemica**
- **Twitter @covid_moonshot**
- **Twitter #BucketListPapers**
- https://www.medchemica.com/bucket-list/

About MedChemica

>10 experience in building A.I. Systems for drug discovery

Roche

Genentech
A Member of the Roche Group

syngenta

blueprint MEDICINES

HEPTARES therapeutics

Alkermes
Patient inspired

REVOLUTION MEDICINES

BenevolentAI

bugworks

Exscientia

CATAPULT
Medicines Discovery

Galápagos

CANCER RESEARCH UK

wellcome

astex pharmaceuticals

ROYAL SOCIETY OF CHEMISTRY

UNICAMP

ICR The Institute of Cancer Research

BLUEBERRY THERAPEUTICS

elixir | insight through communication

Medicines for Malaria Venture

SGC

THE UNIVERSITY of LIVERPOOL

AMR CENTRE | The UK R&D Centre for Antimicrobial Resistance

MANCHESTER 1824
The University of Manchester

Liverpool John Moores University

BigChem

UNIVERSITY OF LEEDS

The University Of Sheffield.

THE UNIVERSITY OF WARWICK

UNIVERSITY OF DUNDEE

- Founded in 2012 by AZ AP Medicinal / Computational chemists to accelerate drug hunting by exploiting data driven knowledge
- Domain leaders in SAR knowledge extraction and knowledge based design
- > 11 years experience of building AI systems that suggest actions to chemists (7 years as MedChemica)
- Creators of largest ever documented database of medicinal chemistry ADMET knowledge

MedChemica Publications

AI Software Platforms

MCPAIRS ENTERPRISE

- Complete In-house platform
- Analysis of own data and automated updating
- Design tool access to all chemists
- Custom fitting (Software-as-a-Service)

Medium to large pharma, agrochemical and materials research

MCPAIRS ONLINE

- Secure web-based AI design platform
- CHEMBL, Patent data analysed
- Merged into one knowledgebase

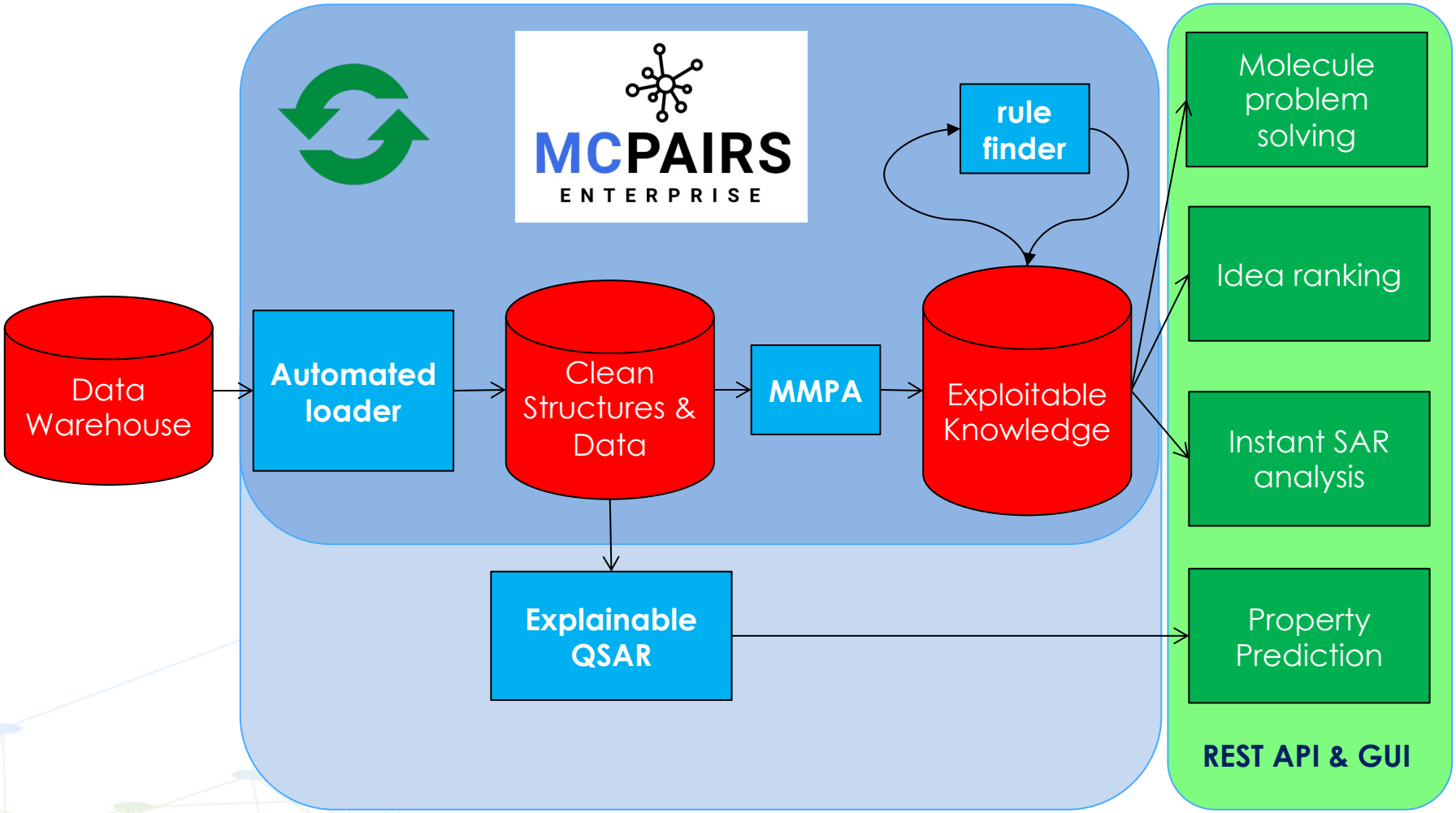
One stop GUI
Design tool
Biotech, Universities and Foundations

AI systems that enhance Chemists

- Principles:
  - Evaluate with directly observable data
  - Expose conflicting views
  - Continuous learning and improvement
  - Place in context

- Translating into practice:
  - Methods that directly connect to chemical structures & data
  - Make all methods auditable
  - Automate updates and track metrics
  - Integrate automated systems and chemists ideas

Explainable AI for Medicinal Chemistry Design



Simple submission and control of the Goal

MedChemica
CREATING A STEP CHANGE IN MEDICINAL CHEMISTRY

Rule Design | Home | Log Out | Modules

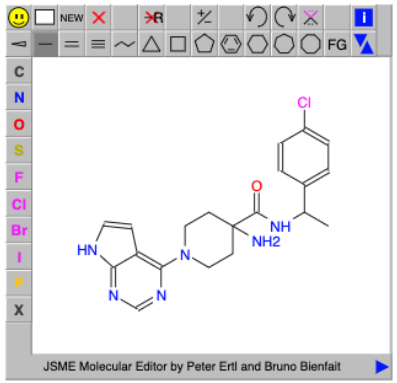
Input Molecule*

Compound Name:

chembl2325997

SMILES*:

CC(c1ccc(cc1)Cl)NC(=O)C2(CCN(CC2)c3c4cc[nH]c4ncn3)N



Sub-structure Lock

SMARTS:

CC(NC(=O)C3(N)CCN(c1n[c;H1]nc2[nH][c;H1][c;H

Draw

Goal*

Direction*:

Increase

Endpoint*:

solubility

Select

Advanced Filters

- [ ] **Molecular charge:**
- [ ] **HBA:**
- [ ] **HBD:**
- [ ] **CLogP:**
- [ ] **RMM:**
- [ ] **PSA:**
- [x] **Specificity:** [1 - 4]

Submit

	#	Timestamp	Status	Number of Products	
❯	0	Jan 28, 2020, 1:32:48 PM	Complete	1	Save
❯	1	Jan 28, 2020, 1:33:27 PM	Complete	139	Save

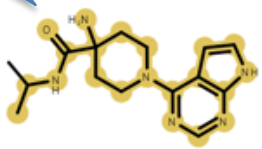
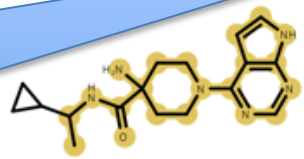
Looking at the results

Results sorted in increasing RMM (Mol Weight)

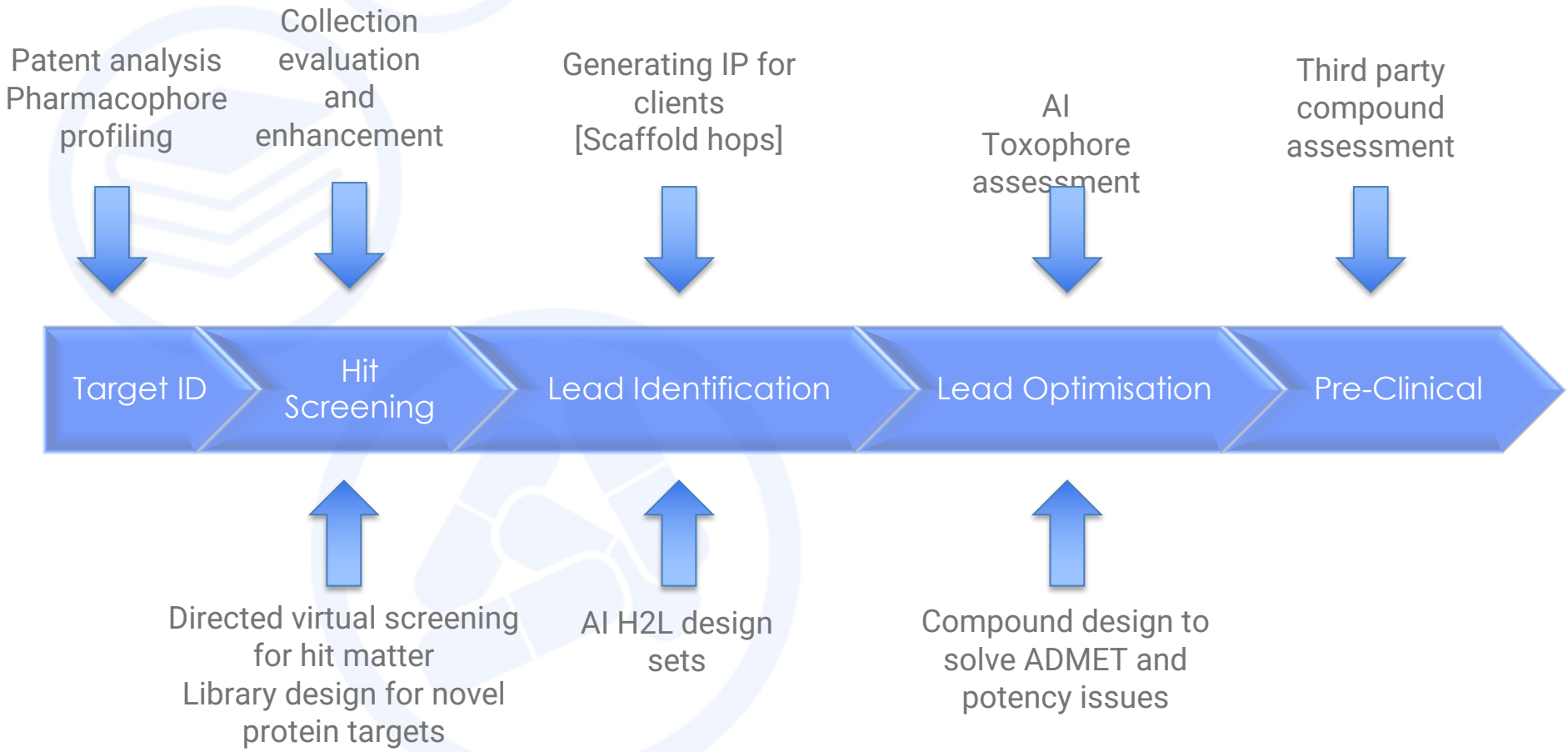
One column per assay
– colour and direction
- LogD decrease, Sol increase

Yellow highlight is the overlap with the input compound

Hyperlink to "Drill back" to the original data

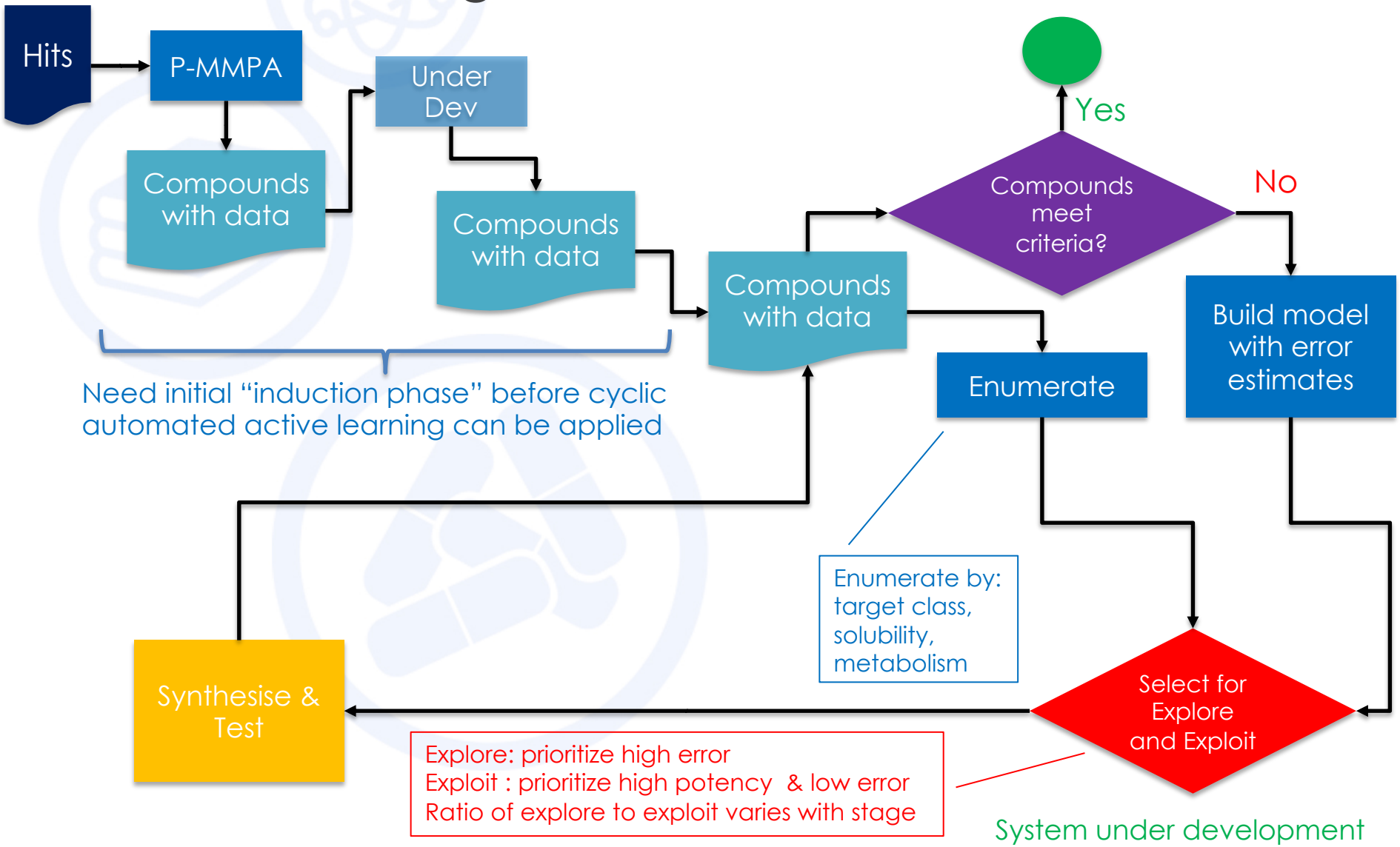
	H	I	J	K	L	M		AD	AT	BJ	
iction	pair data	CLogP	HBA	HBD	PSA	RMM	LogD TM direction	Aq Solubility pH 7.4 [CHEMBL2362975] direction	Aq Solubility pH 7.4 [CHEMBL612558] direction	Aq Solubility comb patent data log(M) direction	PP rat log on Fr dir
 .(CCN(CC1)c2c3cc[nH]c3ncn2)N	matched pair	0.4	7	3	100	302.4	decrease	.	.	increase	.
 .(CCN(CC1)c2c3cc[nH]c3ncn2)N	matched pair	0.7	7	3	100	328.4	NED	.	.	increase	.

Science As A Service (SaaS)



Bespoke Advanced Analytics and Computational Chemistry services through-out the research phase

Active Learning v2



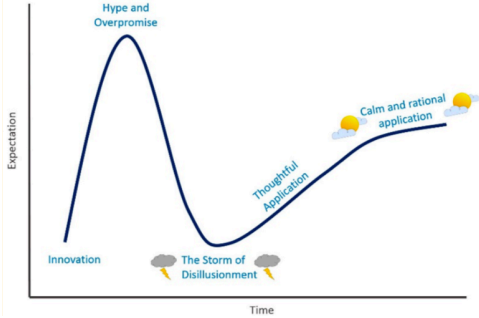
Project examples

An illustration of how MCPairs saves time and money
"Leap-frogging to the best the molecule"

Resistance is ~~Futile~~ Rational

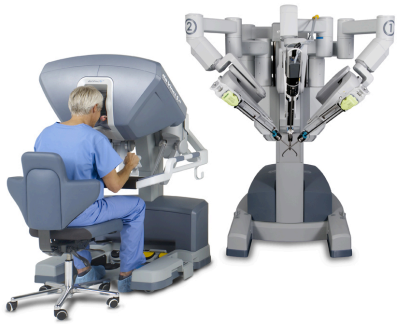
- Is all this AI stuff going to be just another spin round the hype cycle?
  - Artificial Intelligence in Drug Design -The Storm Before the Calm?

    Allan M. Jordan ACS Med. Chem. Lett. 2018, 9, 12, 1150-1152



- What's our defense reaction?
  - Avoid embarrassment
  - Unilateral control – 'I'm in charge'
  - Win-lose conflict framing – 'if I'm better - that is rubbish'
  - Emphasize rationality
  - Avoid inquiry – 'nothing to learn from the past– move on'

- How to Engage Constructively?
  - Define common goals
  - Evaluate with directly observable data
  - Expose conflicting views
  - Continuous learning and improvement
  - Place in context



Argyris C. Organizational traps: leadership, culture, organizational design. Oxford University Press; 2010.
Schein EH. Organizational culture and leadership, Chapter 10: How Leaders Embed and Transmit Culture. 5th Edition. Wiley; 2017.

Project Example 1 – GPR 119 anatagonists

Changing view of SO_2CH_3 on solubility

Leach *et al* 2006

ArH→ArSO_2CH_3 ↑ solubility in line with ↓ logP

Leach *et al* 2012

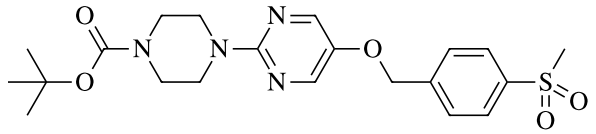
SO_2CH_3 was found to contribute to tight crystal packing via small molecule single crystal x-ray structure determination

5 years of combined data and the application of MMPA

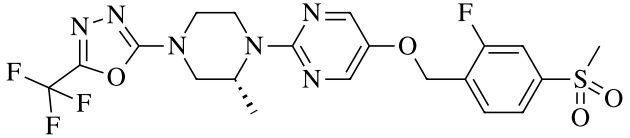
ArH→ArSO_2CH_3 ↓ solubility with ↓ logP

↓ Initial dataset was too small(28 pairs) and unrepresentative and did not include environment

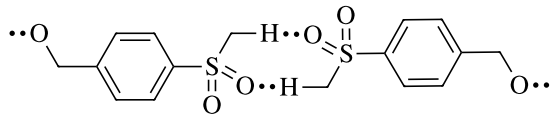
GPR119 - Project

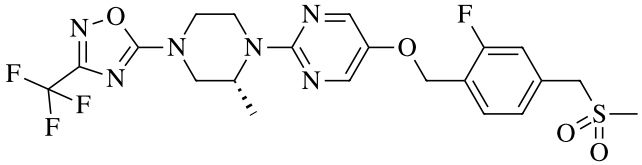


pEC50 7.2 (83%); Sol 0.03 µM



pEC50 7.6 (121%); Sol 1.0 µM; hERG 7 µM



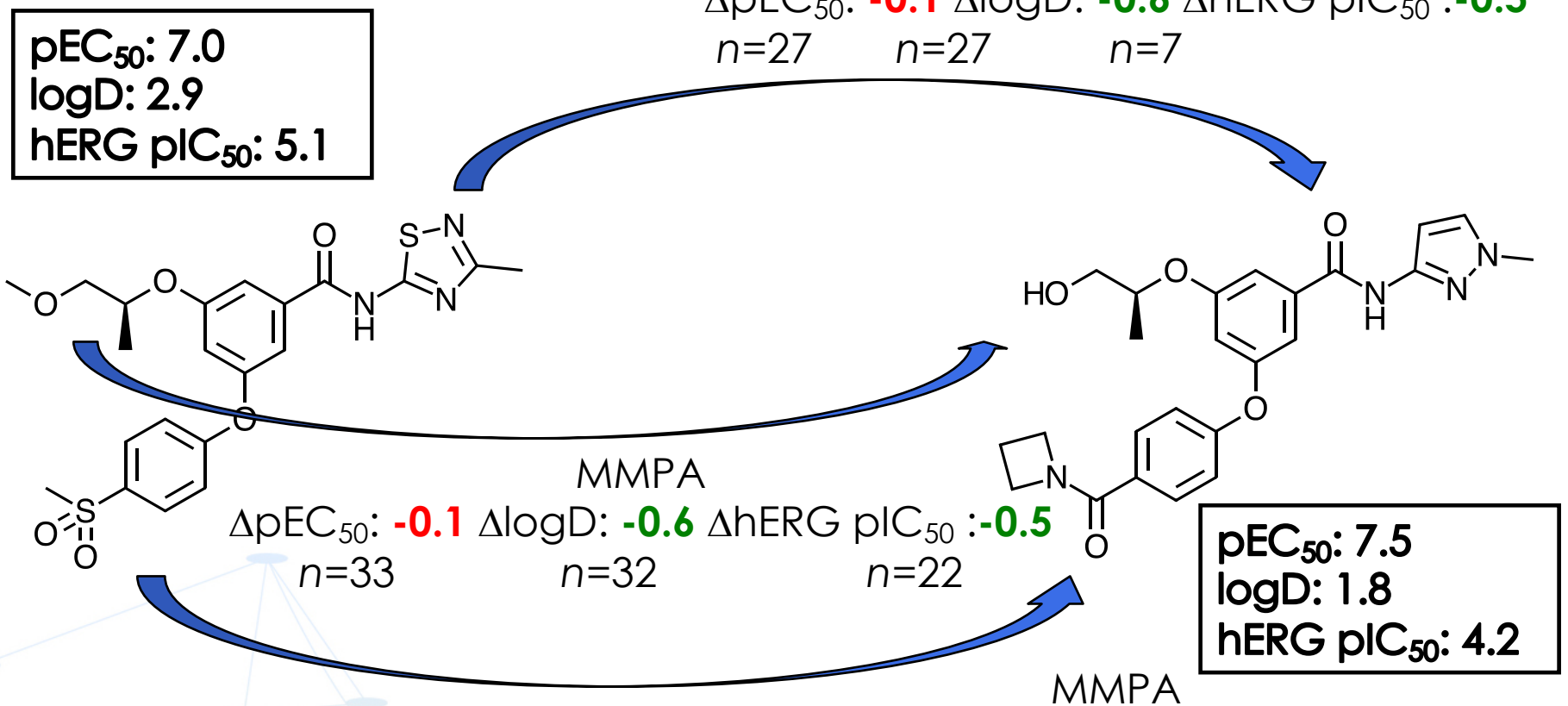


pEC50 8.1 (84%); Sol 1.8 µM; hERG >33 µM

Scott, J.; Leach, A.G.; *et al Med. Chem. Commun.*, **2013**, 4, 95-100
Scott, J.; Leach, A.G.; *et al J.Med. Chem.* **2012**, 55, 5361-5379.
Oxadiazoles in medicinal chemistry. Boström, J. *et al*. (2012) J. Med. Chem. 55, 1817–1830

Project example 2 - Glucokinase Activators

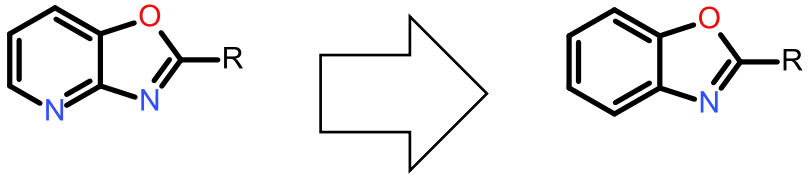
- Fix hERG problem whilst maintaining potency



Waring *et al,* Med. Chem. Commun., **2011**, 2, 775

Project 7 - A Less Simple Example
Increase logD and gain solubility

Question:

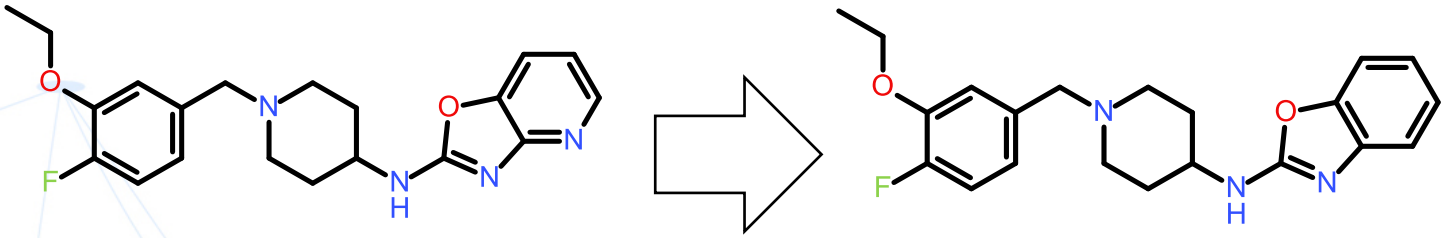


What is the effect on lipophilicity and solubility? Roche data is inconclusive! (2 pairs for logD, 1 pair for solubility)

Available Statistics:

Property	Number of Observations	Direction	Mean Change	Probability
logD	8	Increase	1.2	100%
Log(Solubility)	14	Increase	1.4	92%

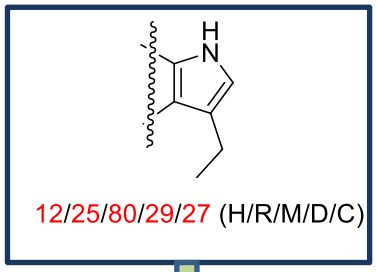
Roche Example:



logD = 2.65
Kinetic solubility = 84 µg/ml
IC50 SST5 = 0.8 µM

logD = 3.63
Kinetic solubility = >452 µg/ml
IC50 SST5 = 0.19 µM

Project 8 - Base of Success Story from Genentech



12/25/80/29/27 (H/R/M/D/C)

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

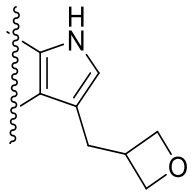
Enumeration

193 compounds Enumerated

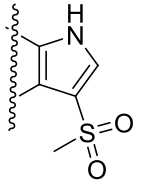
Calculated Property Docking

8 compounds synthesized

Objective: improve metabolic stability



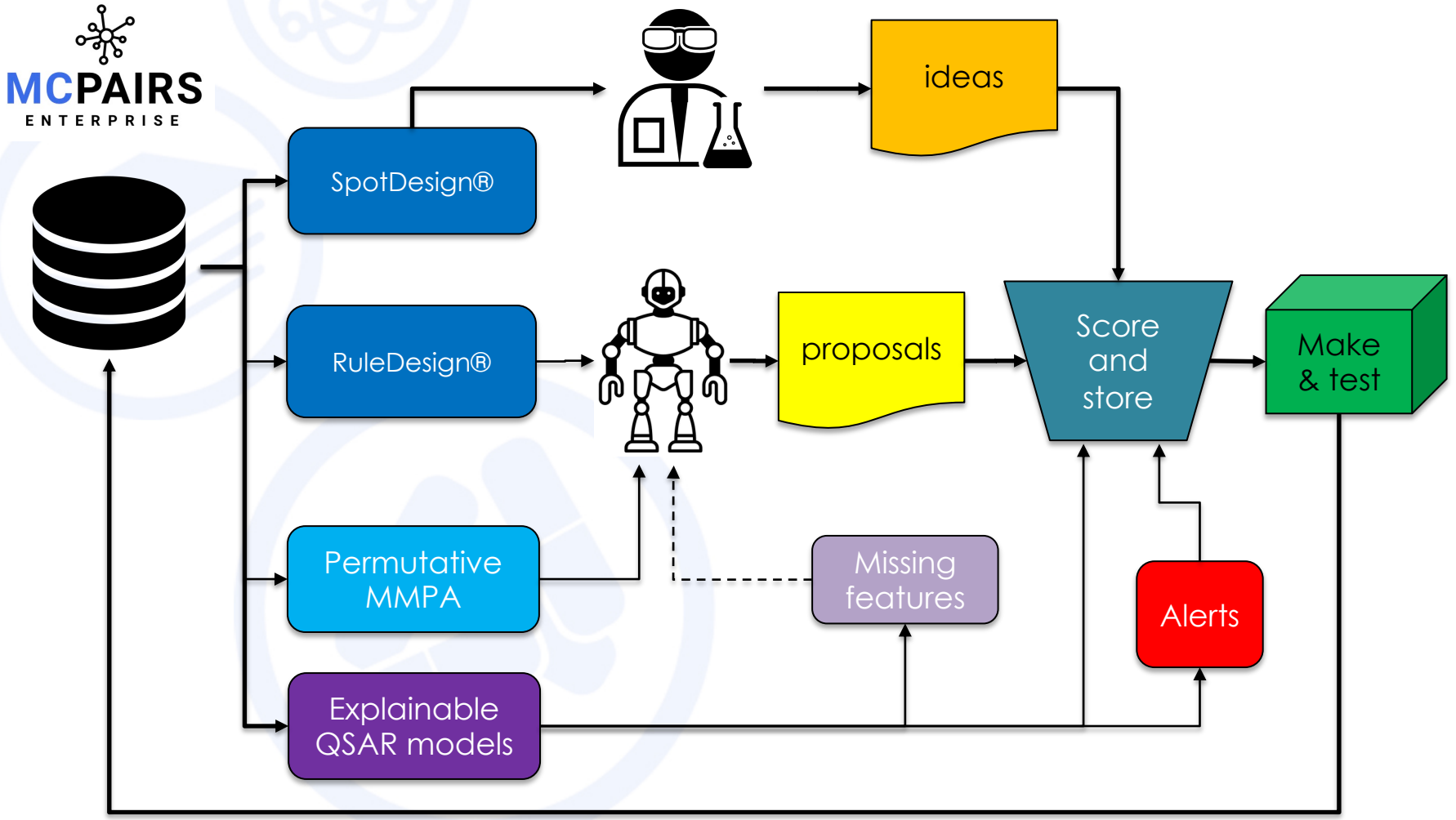
3/15/52/21/33 (H/R/M/D/C)



4/5/2/16/15 (H/R/M/D/C)

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

Augmented Chemists – Data is everything!

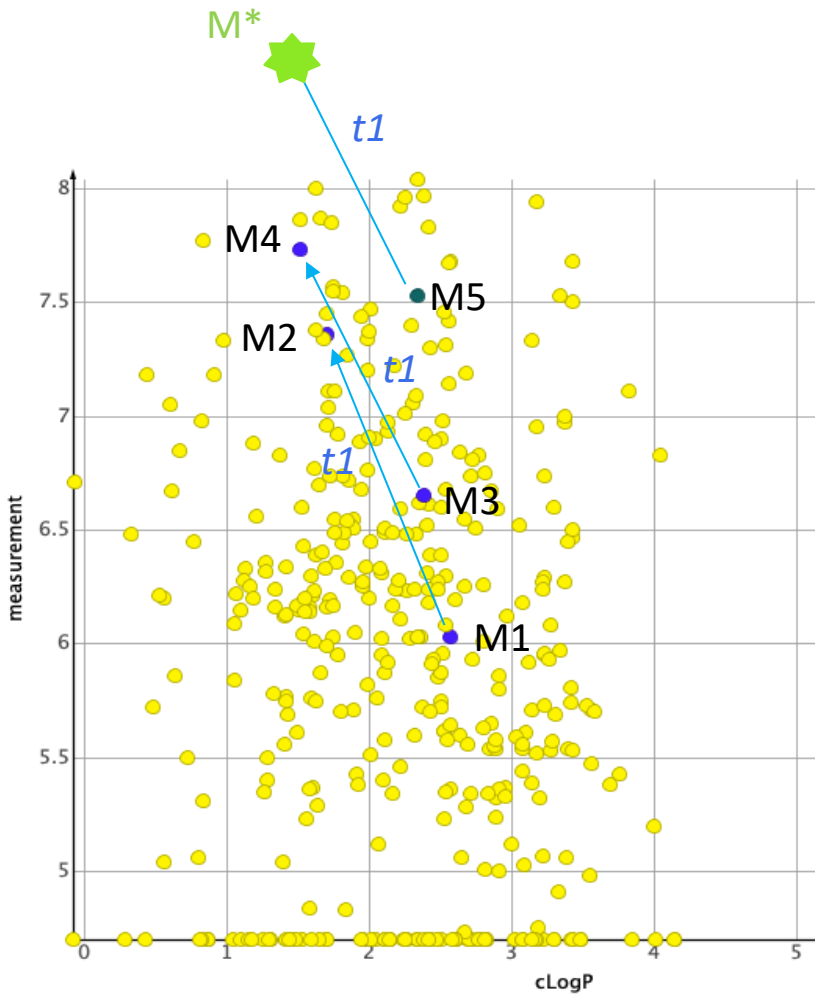


Missing features

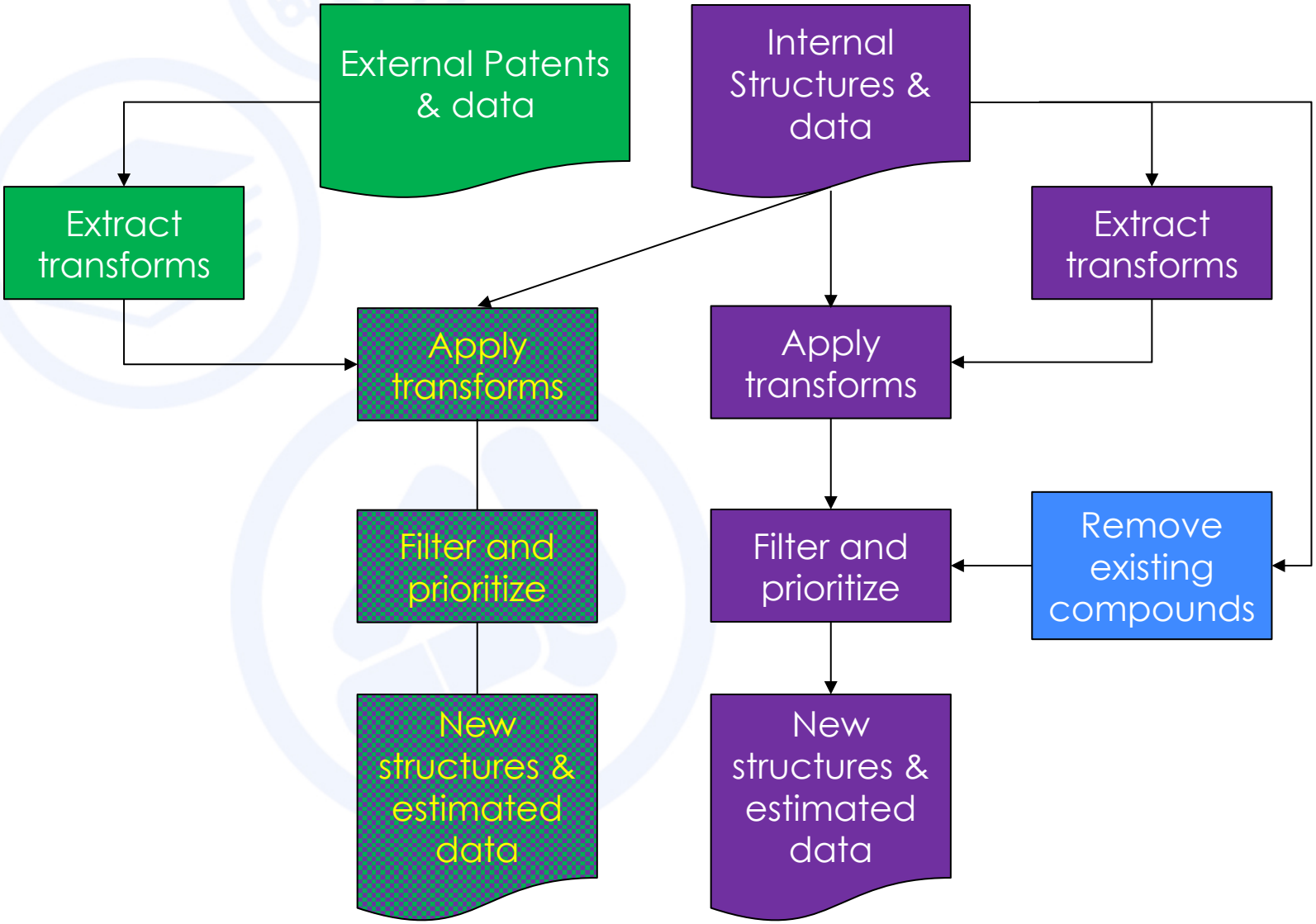
The company that invests in data is the going to be the company that wins in AI

Permutative MMPA (like Free-Wilson)

- **Level 1 - Augmentation**
- Find all matched pairs of a given dataset & extract ΔpIC50 and the transforms between them
- Aggregate transformations with median ΔpIC50 and count of pairs
- Apply all transformations back to the initial compounds in the data set (at the most specific environment level) ***NO R GROUP MAPPING REQUIRED !!!***
- Predicted pIC50 = substrate pIC50 + median ΔpIC50
- Remove existing compounds
- Prioritize new compounds by pIC50 estimate



Exploit Internal and Patent Data



Client Oncology PPI project example

- 386 patent compounds analyzed
- 6024 pair relationships found(39% - good number of MMPs)

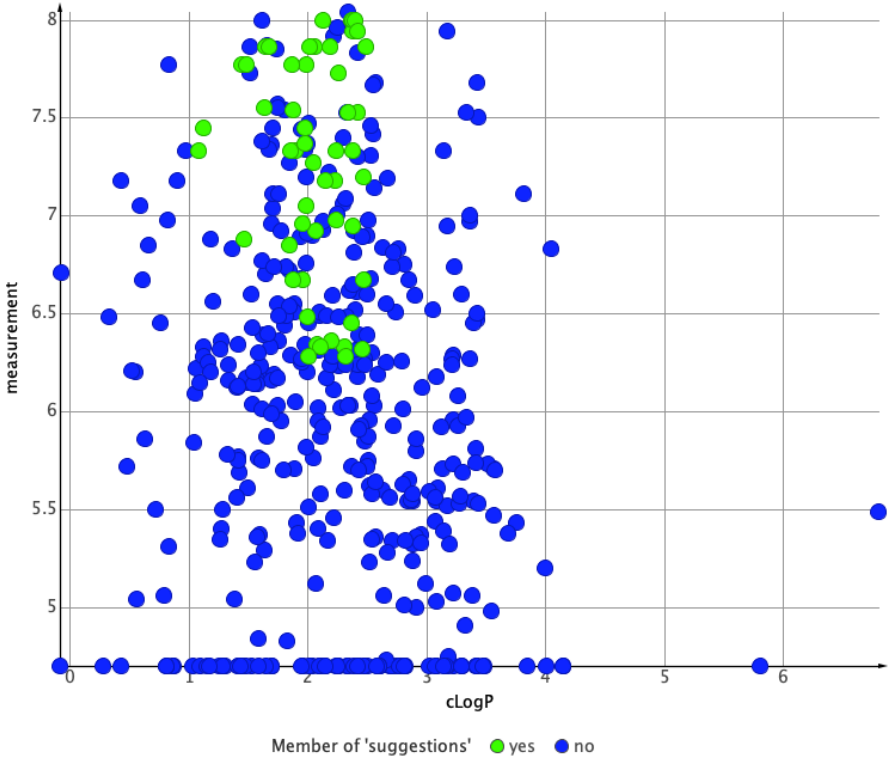
- Permutative MMPA process:
- Apply to own series,
- Then filter:
  - remove undesirable substructure
  - Estimated potency >= 6.5, clogP <= 2.5

- 52 suggestions

Explainable
- Visible, original real world compounds and measurement

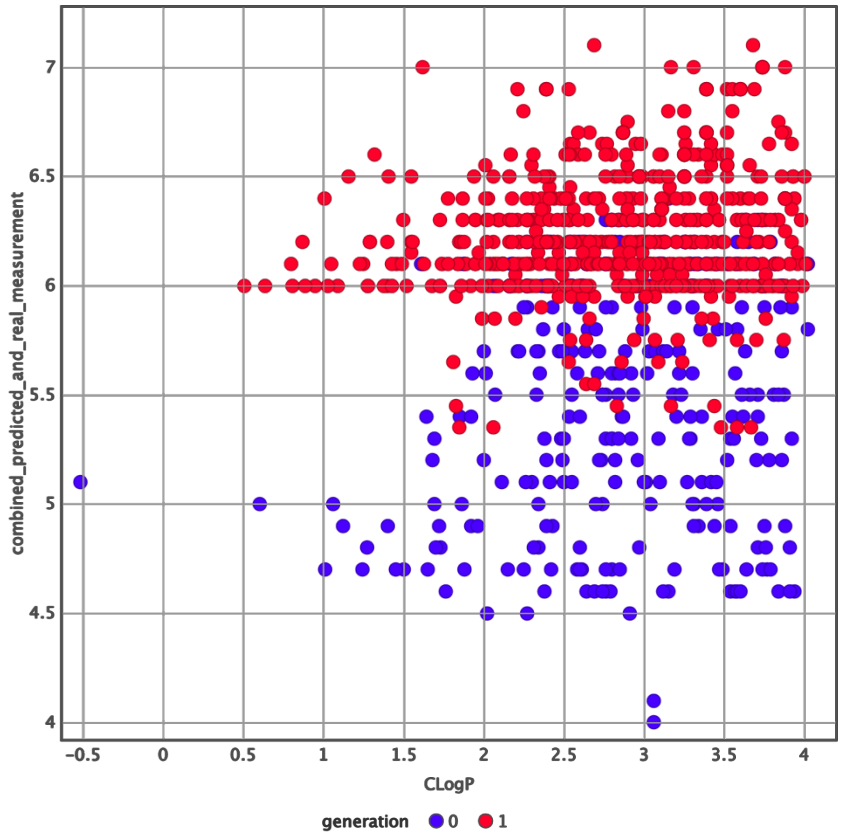
Actionable
- Prioritises 'realistic' next step compounds.

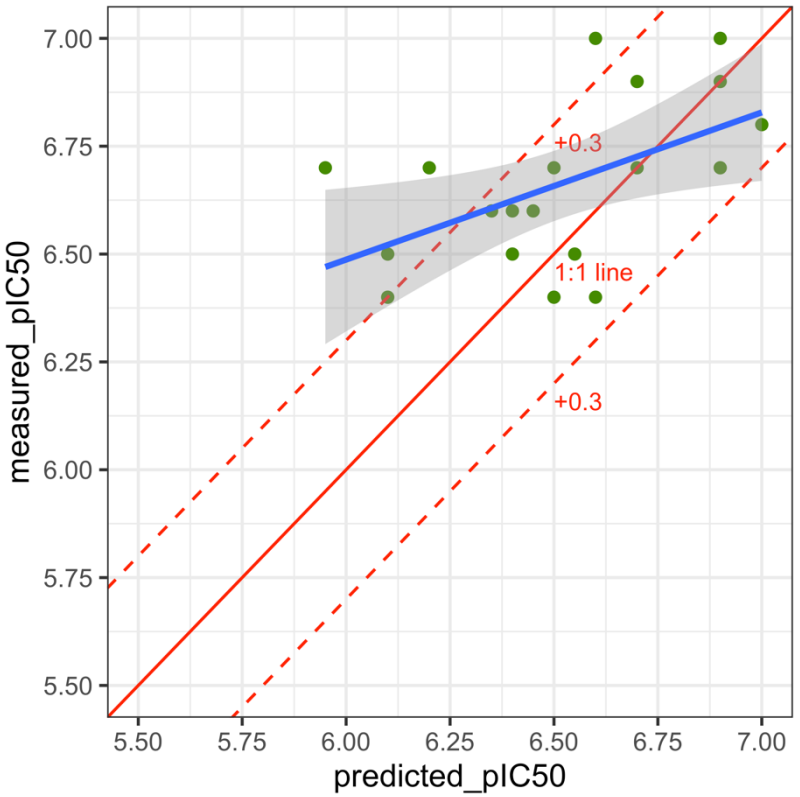
Measurement =
p(TR-FRET nucleotide exchange assay pIC50) or estimated pIC50 from seed value + ΔpIC50



Covid Moonshot - pMMPA example

- 313 Compounds from Covid Moonshot lead series all with pIC50 <6.3 (500nM) fed into pMMPA process,
- New compounds are shown in red, compounds from the input data set in blue
- 671 suggestions(clogP <=4, RMM <=550), 18 with predicted pIC50 > 6 synthesised

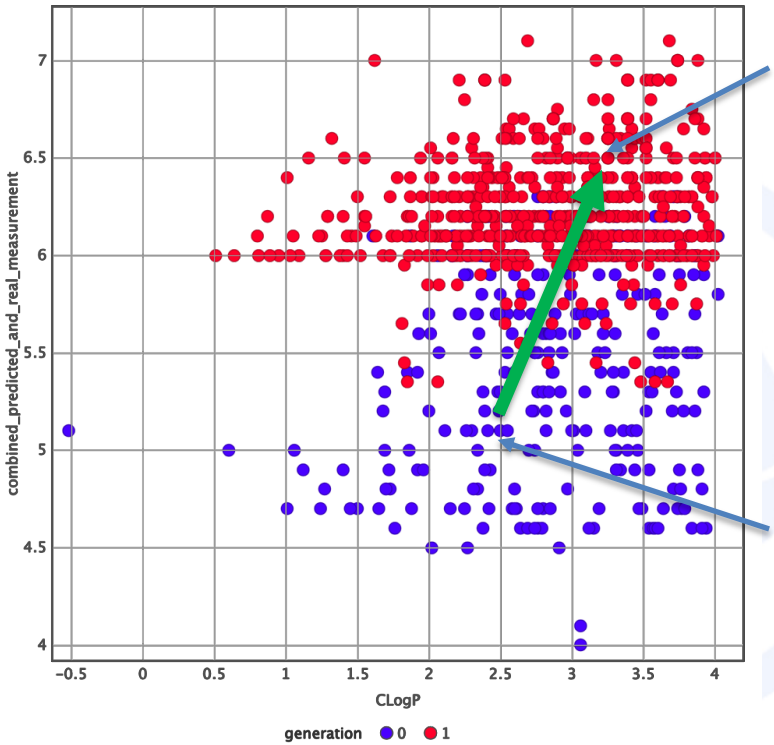


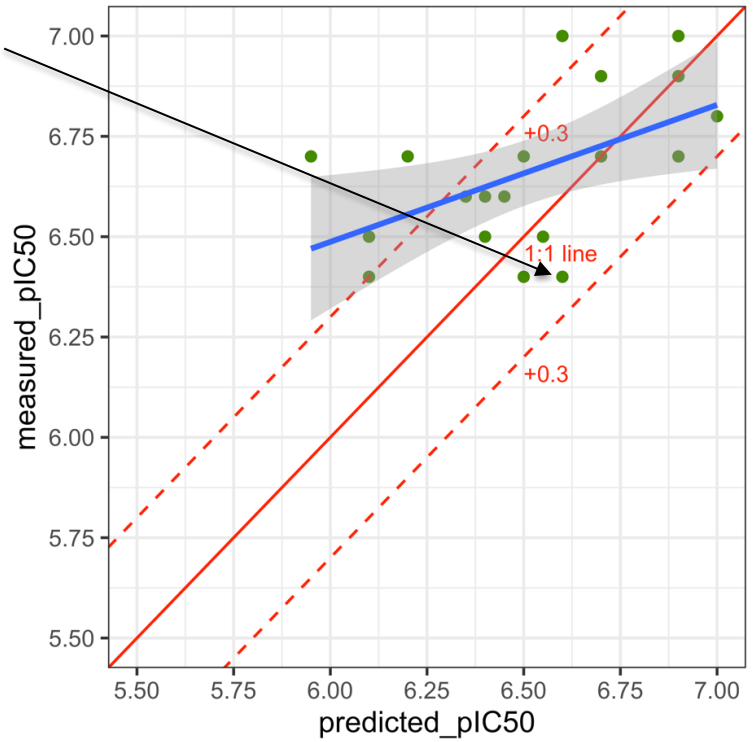


Griffen, E.J.; Full Covid Moonshot talk - Cambridge Med Chem – Sept 2021

Covid Moonshot - pMMPA example

- 313 Compounds from Covid Moonshot lead series all with pIC50 <6.3 (500nM) fed into pMMPA process,
- Compounds from the input data set in blue, new compounds are shown in red,
- 671 suggestions(clogP <=4, RMM <=550), 18 with predicted pIC50 > 6 synthesised





Griffen, E.J.; Full Covid Moonshot talk - Cambridge Med Chem – Sept 2021