UNC ESHELMAN SCHOOL OF PHARMACY

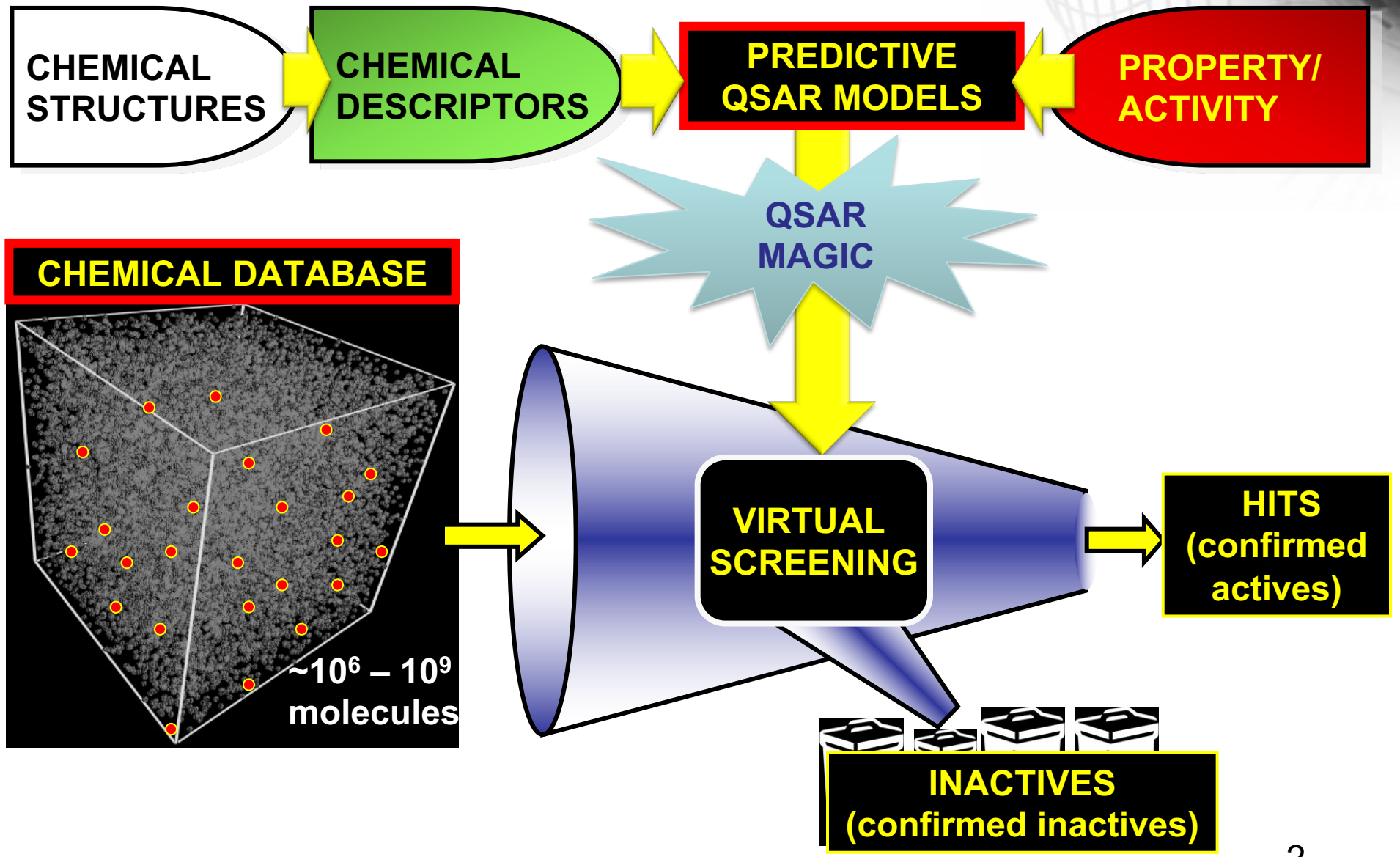
MML UNC.EDU

Rigor and reproducibility of Cheminformatics models: from data curation to the experimental validation

Alexander Tropsha

University of North Carolina, Chapel Hill, USA

The chief utility of CADD: Hit identification in external libraries



Drug discovery in the age of Big Data: need for new methods and careful, automated curation!

National Library of Medicine
National Center for Biotechnology Information

116M Compounds 310M Substances 293M Bioactivities

PubChem About Docs

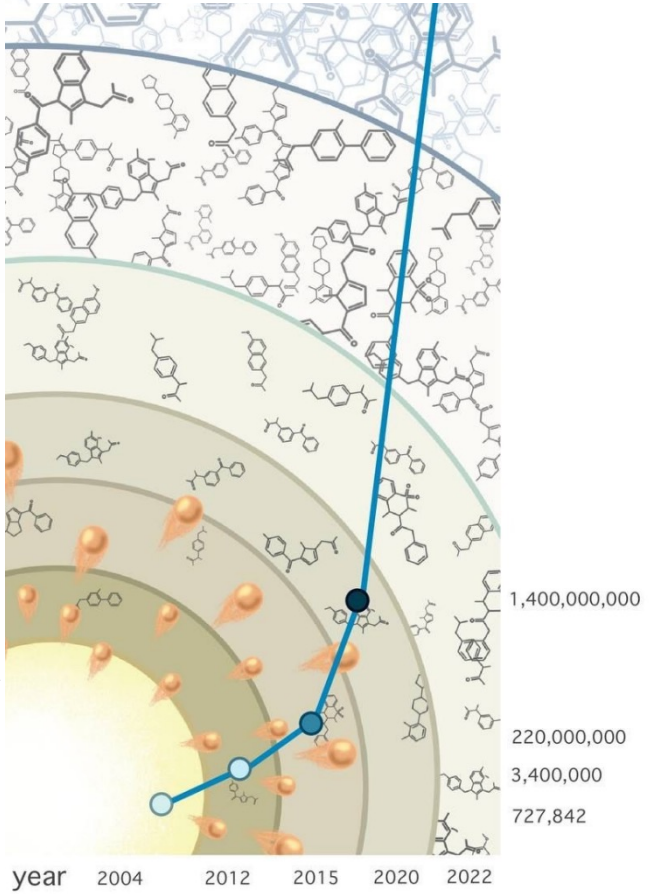
Europe PMC

Over 41 million abstracts and 8.7 million full-text articles, adding over 1.7 million new articles annually.

Knowledge mining

Virtual Screening

40+B of purchasable chemicals!!!



Cherkasov, A. The 'Big Bang' of Chemical Universe. Nature Chemical Biology, 19, 667–668 (2023)

Pandey M., et al. The transformational role of GPU computing and deep learning in drug discovery. Nature Mach. Intel. 2022 4, 211–221

Tropsha. A., et al. Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR. Nature Rev. Drug Disc, 2023, https://doi.org/10.1038/s41573-023-00832-0

Typical elements of QS[A,P,T]R modeling: issues at every step

- Experimental Data = pain
  - Structure
  - Activity
- Model Validation
  - Descriptors
  - Statistical/machine learning techniques
- Prediction (i.e., data imputation)
- Experimental confirmation of predictions

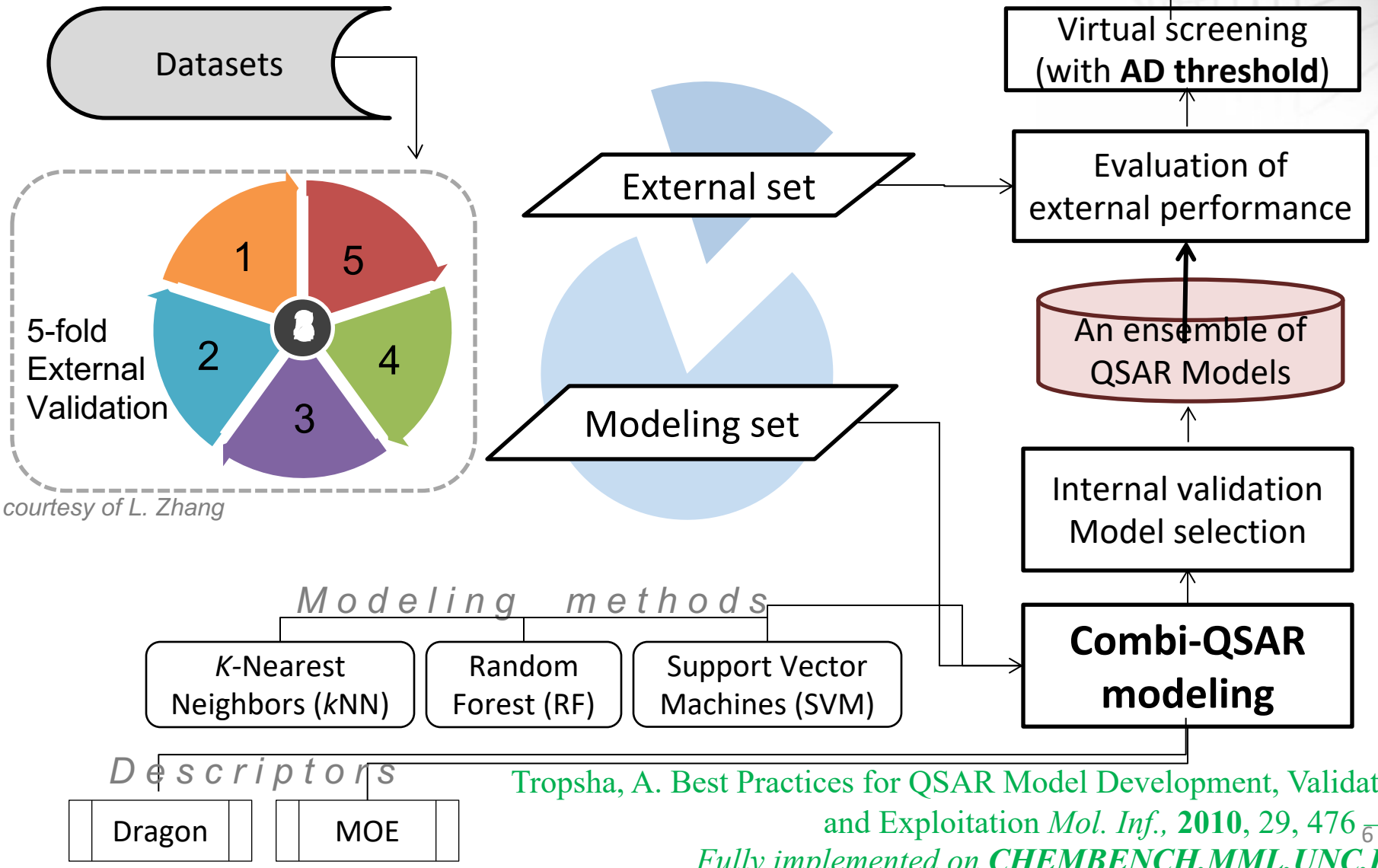
- Reliable models to enable decision support (both in research and for regulatory approval) = gain

Published guidance on model development and validation: The OECD Principles

To facilitate the consideration of a QSAR model for regulatory purposes, it should be associated with the following information:

- a defined endpoint
- an unambiguous algorithm;
- a defined domain of applicability
- appropriate measures of goodness-of-fit, robustness and predictivity
- a mechanistic interpretation, if possible;
- Should be added: data used for modeling should be carefully curated

QSAR Modeling Workflow: the importance of rigorous validation



Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation *Mol. Inf.*, **2010**, 29, 476 – 488

Fully implemented on ***CHEMBENCH.MML.UNC.EDU***

21 “how not to do QSAR” principles

Table 1. Types of error in QSAR/QSPR development and use.

No.	*Type of error*	*Relevant OECD principle(s)*
1	Failure to take account of data heterogeneity	1
2	Use of inappropriate endpoint data	1
3	Use of collinear descriptors	2, 4, 5
4	Use of incomprehensible descriptors	2, 5
5	Error in descriptor values	2
6	Poor transferability of QSAR/QSPR	2
7	Inadequate/undefined applicability domain	3
8	Unacknowledged omission of data points	3
9	Use of inadequate data	3
10	Replication of compounds in dataset	3
11	Too narrow a range of endpoint values	3
12	Over-fitting of data	4
13	Use of excessive numbers of descriptors in a QSAR/QSPR	4
14	Lack of/inadequate statistics	4
15	Incorrect calculation	4
16	Lack of descriptor auto-scaling	4
17	Misuse/misinterpretation of statistics	4
18	No consideration of distribution of residuals	4
19	Inadequate training/test set selection	4
20	Failure to validate a QSAR/QSPR correctly	4
21	Lack of mechanistic interpretation	5

Dearden JC et al., **2009,** SAR and QSAR in Environmental Research, Vol. 20, Nos. 3–4, April–June 2009, 241

Critical assessment of published QSAR models

- Issues
  - Primary data is not curated
  - Correlations are inflated
  - Outliers are abundant
  - Statistical metrics of models are often inadequate
  - Published models are not validated
  - Mechanistic interpretation is often derived from bad models
- Challenge: develop best model development and publishing practices for cheminformatics papers
  - **The ideal bad cheminformatics paper is the one that was not accepted for publication!**

Some reasons why QSAR models may fail

- No external validation
- Incorrect selection of an external test set
- Incorrect division of a dataset into training and test sets
- Incorrect measure of prediction accuracy
- Not all statistical criteria are used to estimate predictive power of a model
- No applicability domain
- Incorrectly defined applicability domain
- No Y-randomization
- Leverage (structure) and activity outliers are not removed
- Modeling set is too small

Some reasons why QSAR models may fail: Misiniterpretation of the Models' Predictive Ability, lack or incorrect external validation

MML UNC.EDU

QSAR Pill

- Johnson, S.R. The Trouble with QSAR (or How I Learned To Stop Worrying and Embrace Fallacy). *J. Chem. Inf. Model.* **2008,** *48,* 25-26:

  "The common practice has been to select the model with the best fitness function score and predict a small group of observations that were withheld at the beginning. All too often, the model development process stops here, or, worse, the validation set is poorly predicted, and models are iteratively tested until one predicts this set of compounds well."

A typical example:

A dataset is divided into a training and test set

Multiple QSAR models with high q^2 values are built using training set

QSAR model with the highest R^2 for the test set is selected

Selected model could have poor predictive ability for other compounds

Additional EXTERNAL EVALUATION SETS are necessary

Some reasons why QSAR models may fail: Incorrect division of a dataset into training and test sets

MML UNC.EDU

QSAR Pill

- **Typical division of a dataset into training and test sets**: random
  - **Undesired outcome**:
    - some compounds of the test set can be out of the applicability domain
    - large activity gaps in the training or test set; activity outliers
- **Requirements for training and test sets**:
  - Compounds with maximum and minimum activities of the dataset should be included into the training set (important for methods that cannot extrapolate activities, e.g., kNN).
  - Large activities gaps are not allowed neither in training nor the test set.
  - Each compound of the test set should be close to at least one compound of the training set.

Some reasons why QSAR models may fail: using incorrect metric to assess classification QSAR accuracy for biased datasets:

QSAR Pill

- **A typical target function (Classification Rate)**:

  CR=N(classified correctly)/N(total)

  A dataset:

  Class 1: 80 compounds; **Class 2**: 20 compounds

  Model: assign all compounds to Class 1.

  Target function: CR=0.8

  The model appears to have high classification accuracy

- **Better target function**:

  CCR (or BA) =0.5x(Sensitivity+Specificity)

  In the above example, CCR = 0.5

- **General formula**:

$$CCR = \frac{1}{K}\sum_{k=1}^{K}\frac{N_k^{corr}}{N_k^{total}}$$

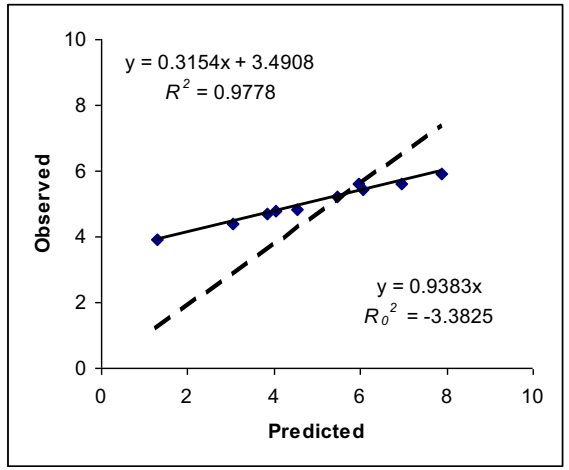
***K* – the number of classes**

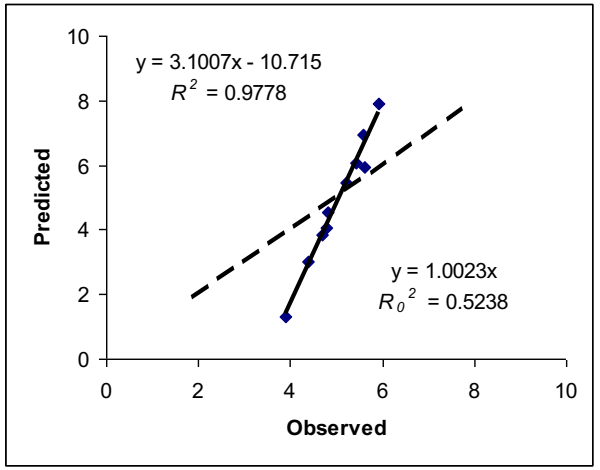
N_k^{corr} – the number of compounds of class *k* assigned to class *k*

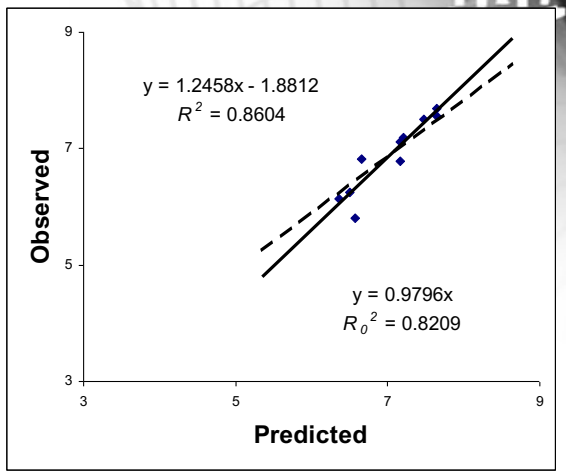
N_k^{total} – total number of compounds of class *k*

- **For categorical response variable, target functions can depend also on the absolute errors (differences between predicted and observed classes).**

How to define predictive accuracy of a QSAR model







Regression

$$\tilde{y}^r = a' y + b'$$

$$a' = \frac{\sum (y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sum (y_i - \bar{y})^2}$$

$$a = \frac{\sum (y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sum (\tilde{y}_i - \bar{\tilde{y}})^2}$$

$$b = \bar{y} - a\bar{\tilde{y}} \qquad b' = \bar{\tilde{y}} - a'\bar{y}$$

$$y^{r_0} = k\tilde{y}$$

$$k = \frac{\sum y_i \tilde{y}_i}{\sum \tilde{y}_i^2}$$

Correlation coefficient

$$R = \frac{\sum (y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\tilde{y}_i - \bar{\tilde{y}})^2}}$$

Regression through the origin

$$\tilde{y}^{r_0} = k' y$$

$$k' = \frac{\sum y_i \tilde{y}_i}{\sum y_i^2}$$

Coefficients of determination

$$R_0^2 = 1 - \frac{\sum (\tilde{y}_i - y_i^{r_0})^2}{\sum (\tilde{y}_i - \bar{\tilde{y}})^2}$$

$$R'^2_0 = 1 - \frac{\sum (y_i - \tilde{y}_i^{r_0})^2}{\sum (y_i - \bar{y})^2}$$

CRITERIA

$$q^2 > 0.5; R^2 > 0.6;$$

$$k \text{ or } k' \approx 1.0; R_0^2 \text{ or } R_0'^2 \approx R^2$$

Some reasons why QSAR models may fail: No Applicability Domain is defined for the Model

MML UNC.EDU

- **Compounds which are highly dissimilar from all compounds of the training set (according to the set of descriptors selected) cannot be predicted reliably**

  Lack of the AD:

  - unjustified extrapolation

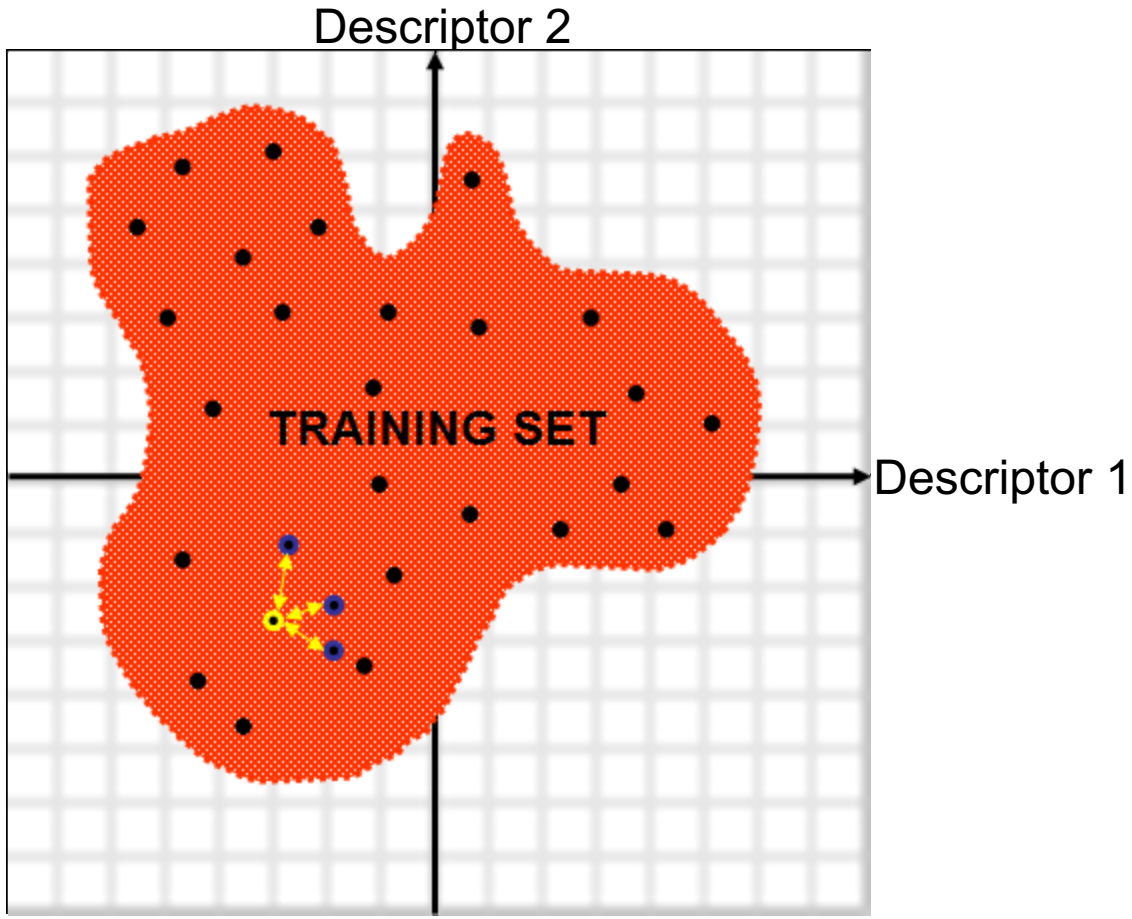
  - wrong prediction

  Typical situation:

  a compound of the test set for which error of prediction is high is considered as outlier

  HOWEVER: a compound of the test set dissimilar from all compounds of the training set can be by chance predicted accurately

Applicability domain of QSAR models

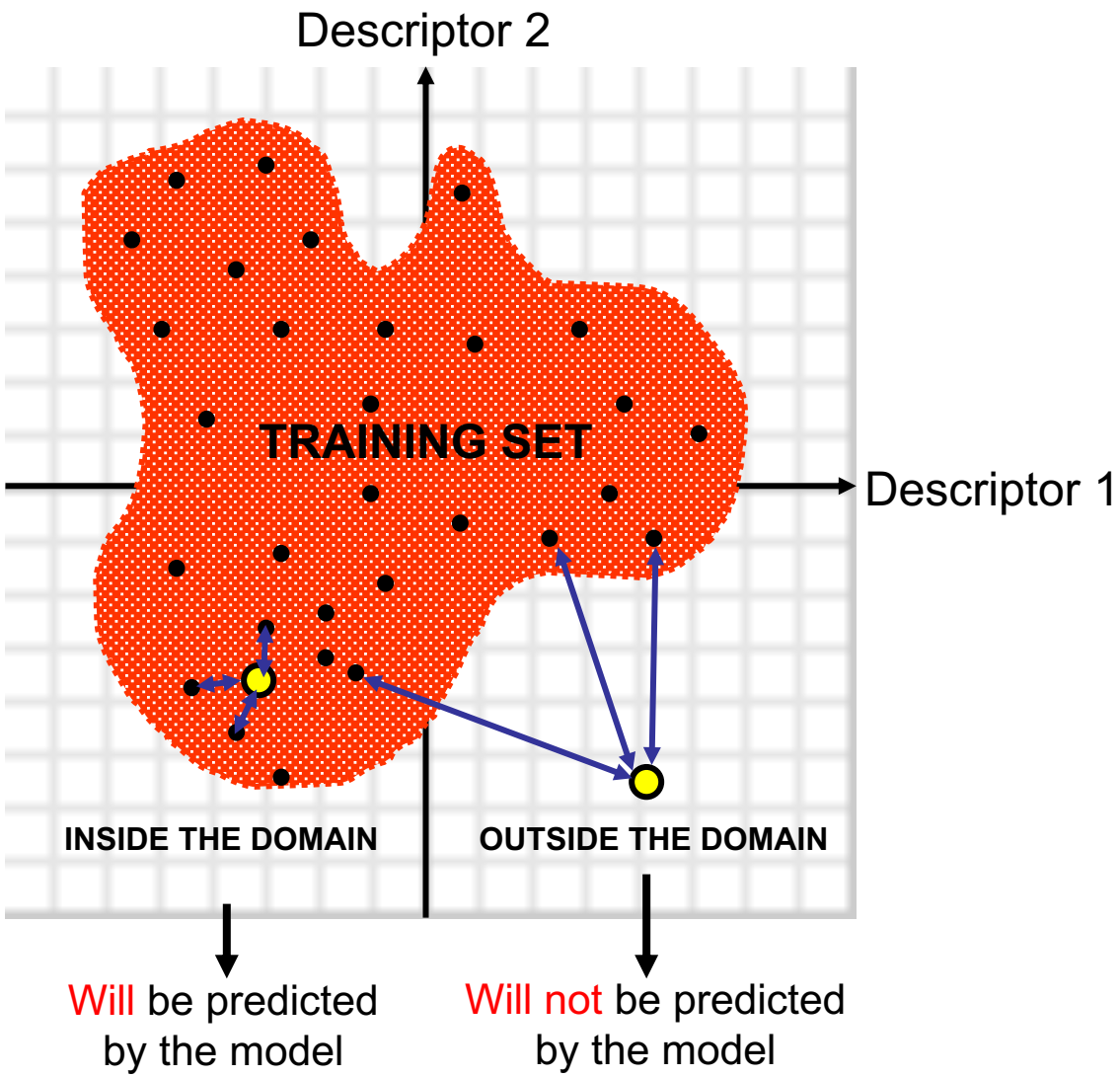


For a given model, two parameters are calculated:

- **$<D_k>$** : average Euclidian distance between each compound of the training set and its *k* nearest neighbors in the descriptors space;

- **s_k** : standard deviation of the distances between each compound of the training set and its k nearest neighbors in the descriptors space.

Applicability domain of QSAR models



For a given model, two parameters are calculated:

- **$<D_k>$** : average euclidian distance between each compound of the training set and its *k* nearest neighbors in the descriptors space;
- **s_k** : standard deviation of the distances between each compound of the training set and its k nearest neighbors in the descriptors space.

= NEW COMPOUND

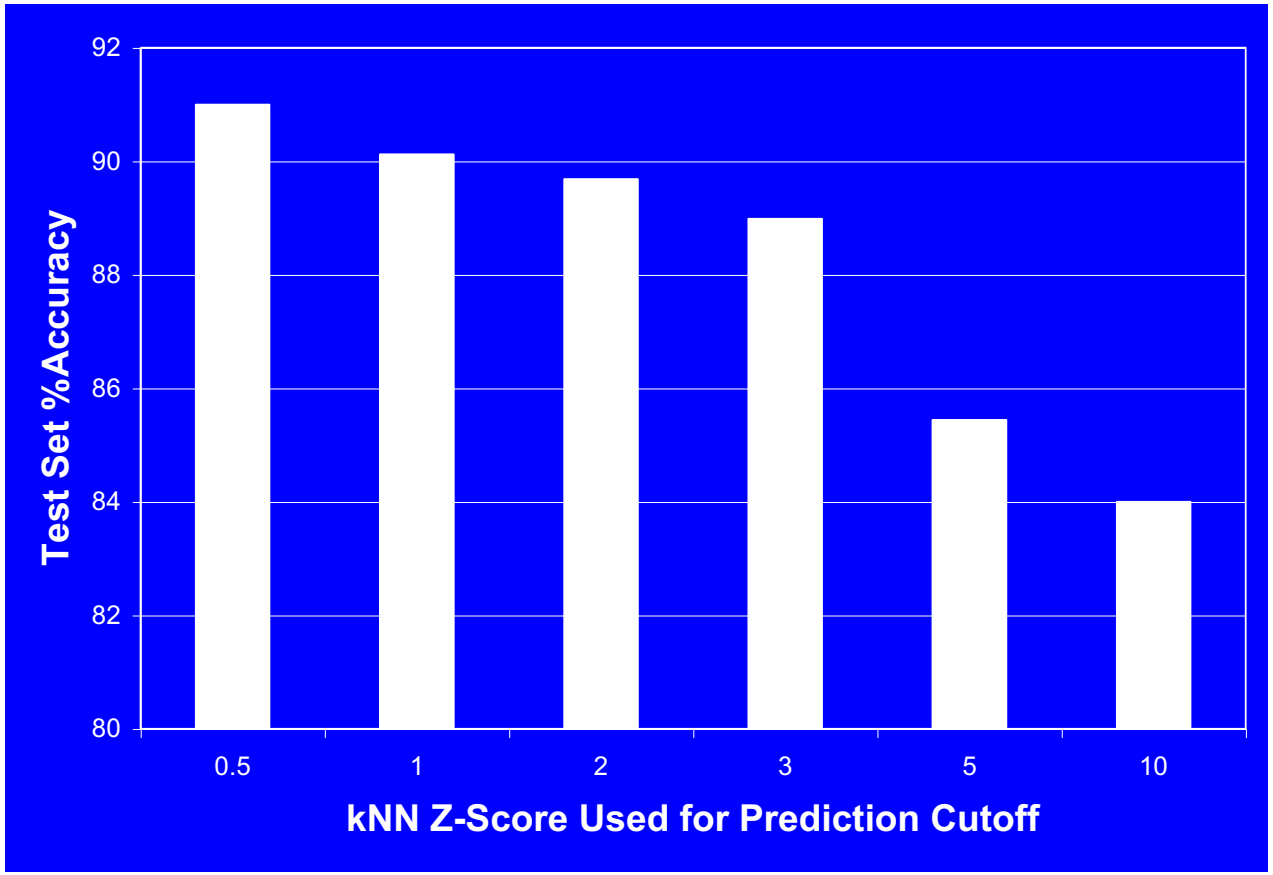
For each test compound *i*, the distance D_i is calculated as the average of the distances between i and its *k* nearest neighbors in the training set.

The new compound will be predicted by the model, only if :

$$D_i \leq <D_k> + Z \times s_k$$

with Z, an empirical parameter (0.5 by default)

Applicability domain vs. prediction accuracy (Ames Genotoxicity dataset)



Some reasons why QSAR models may fail: Y-randomization test is not carried out

MML UNC.EDU

QSAR Pill

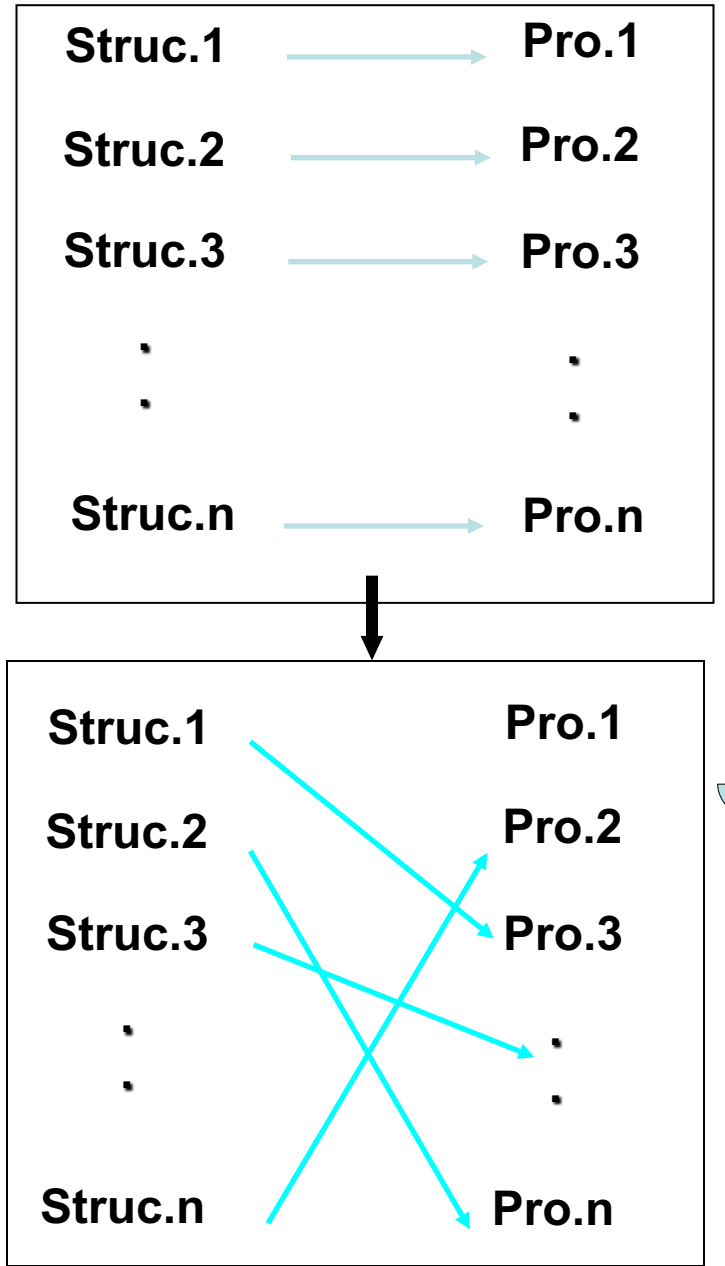
- **Y-randomization test**:
  - Scramble activities of the training set
  - Build models and get model statistics.
  - If statistics are comparable to those obtained for models built with real activities of the training set, the last are unreliable and should be discarded.

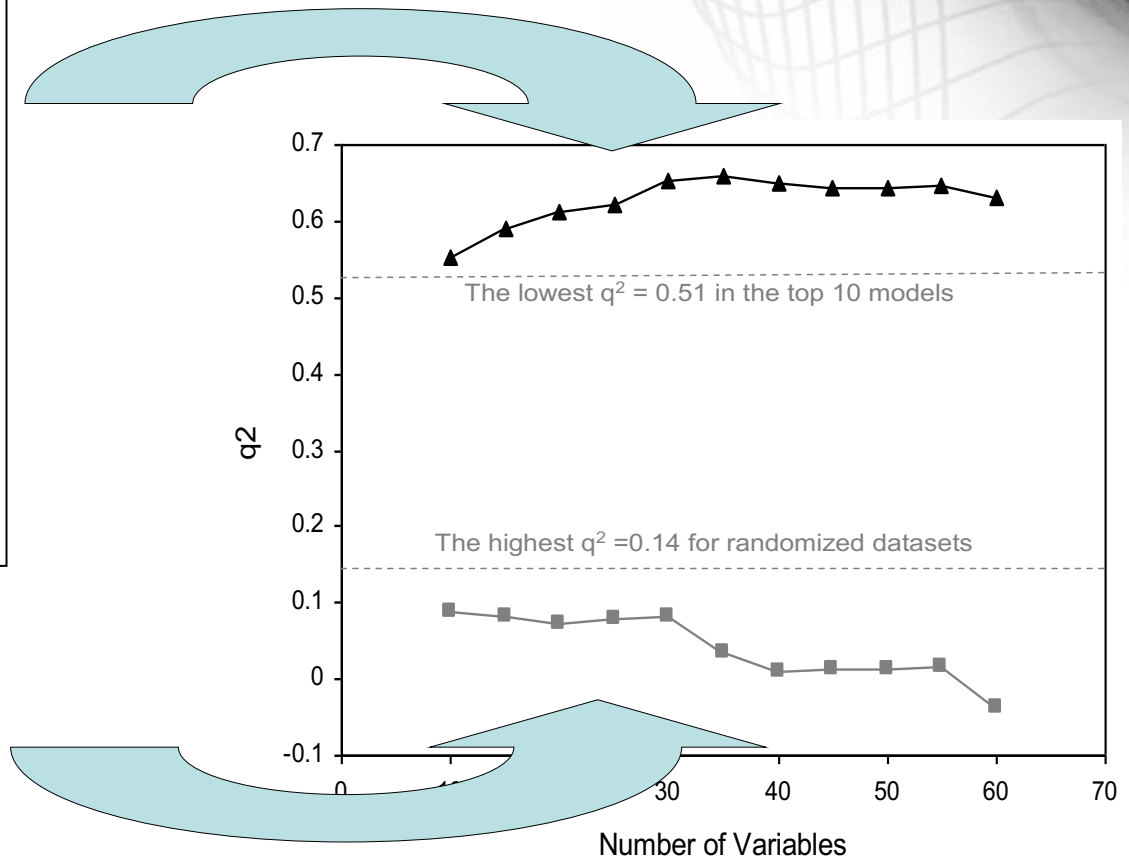
Frequently, Y-randomization test is not carried out.

Y-randomization test is of particular importance, if there is:

- a small number of compounds in the training or test set
- the response variable is categorical

Activity randomization: model robustness





Training set with real property values is expected to produce much higher q^2 values than the same set with randomized property values.

Detection and removal of outliers

MML UNC.EDU

QSAR Pill

- Many potential outliers can be detected in the dataset prior to QSAR studies, but typically this is not done.

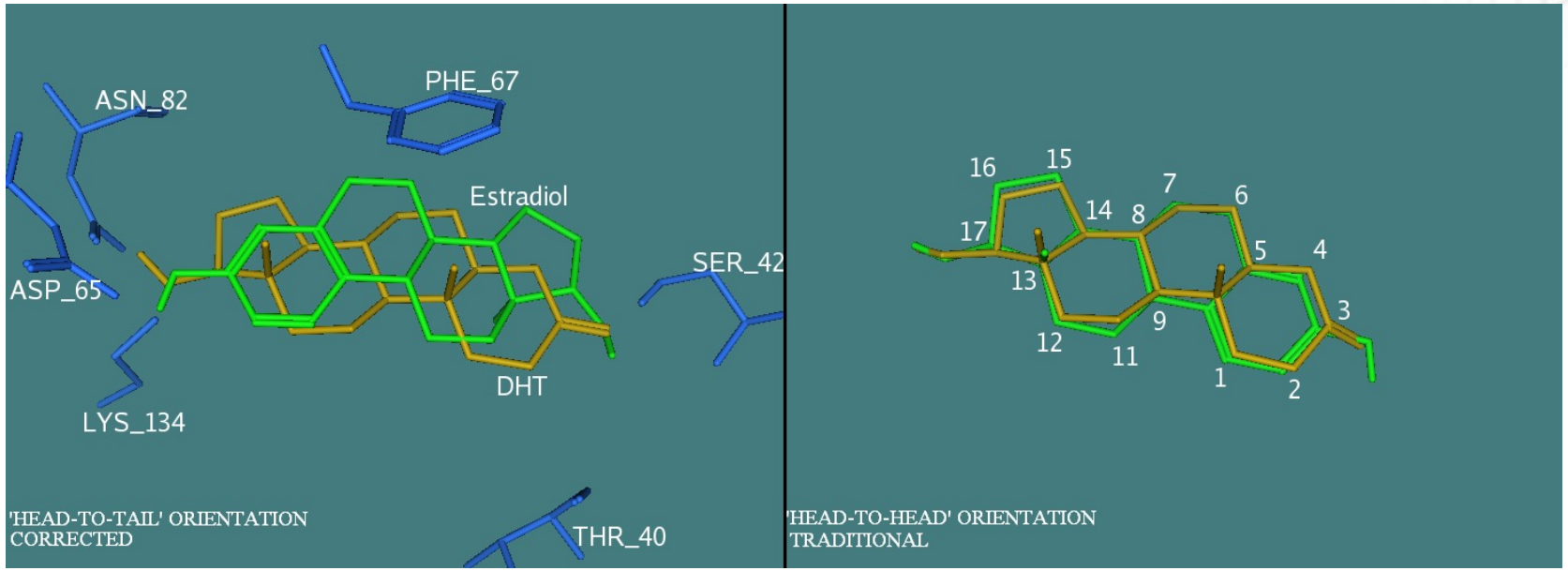
- **Two types of outliers**

  - Leverage outliers: compounds dissimilar from all other compounds in a dataset in the chemistry space.

  -Activity outliers: compounds similar to some other compounds in the dataset, but with activities quite different from those of their nearest neighbors.

Why QSAR models may fail: insensitive descriptors.

Identical q^2 (CoMFA*) of 0.53



Optimal Traditional

Orientations of androgen (DHT shown in gold) and estrogen (estradiol shown in green) within human SHBG steroid-binding site

A. Cherkasov, *JMC*, 2008

Why QSAR models may fail: incorrect structures

- "Slight errors in chemical structures, such as misplacing a Cl atom or swapping hydroxy and methoxy functional groups on a multiple ring structure, can result in significant differences in the accuracy of the prediction for those chemicals.

  Young et al, Are the Chemical Structures in Your QSAR Correct? *QSAR Comb. Sci*. 27, 2008, No. 11-12, 1337 – 1345

- Data Curation
  - Removal of inorganics, salts, and mixtures
  - Aromatization and 2D cleaning
  - Normalization of carboxylic, nitro, etc. groups
  - Elimination of duplicates
  - Standardization of functional group representation
  - Manual cleaning
  - … and then, look at 'em again!

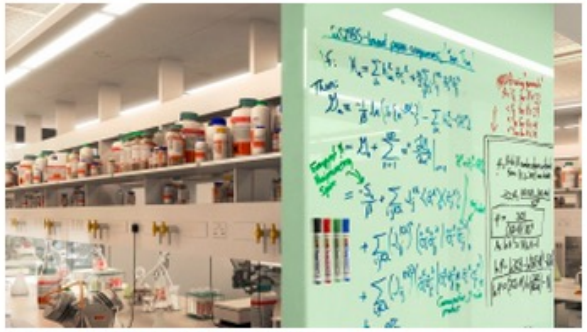
QSAR Pill

TheScientist
EXPLORING LIFE, INSPIRING INNOVATION

Dealing with Irreproducibility

Researchers discuss the growing pressures that are driving increases in retraction rates at AACR.

By Jef Akst | April 8, 2014



FLICKR, UNIVERSITY OF EXETER

Recent years have seen increasing numbers of retractions, higher rates of misconduct and fraud, and general problems of data irreproducibility, spurring the National Institutes of Health (NIH) and others to launch initiatives to improve the quality of research results. Yesterday (April 7), at this year's American Association for Cancer Research (AACR) meeting, researchers gathered in San Diego, California, to discuss why these problems to come to a head—and how to fix them.

"We really have to change our culture and that will not be easy," said Lee Ellis from the University of

be_right_c

Aut

66

ation »

ws

team

3250

0

y call

as

of

ed

of

nBark

tigates

u Sezen's

at Columbia.

pond to

as unlabeled

ompound in

osted in full

ted

DOI: 10.1

when we receive it.

In the last ten years, public online databases have

In the last decade numerous attempts have been made to

Data dependency and data quality are critical issues in QSAR

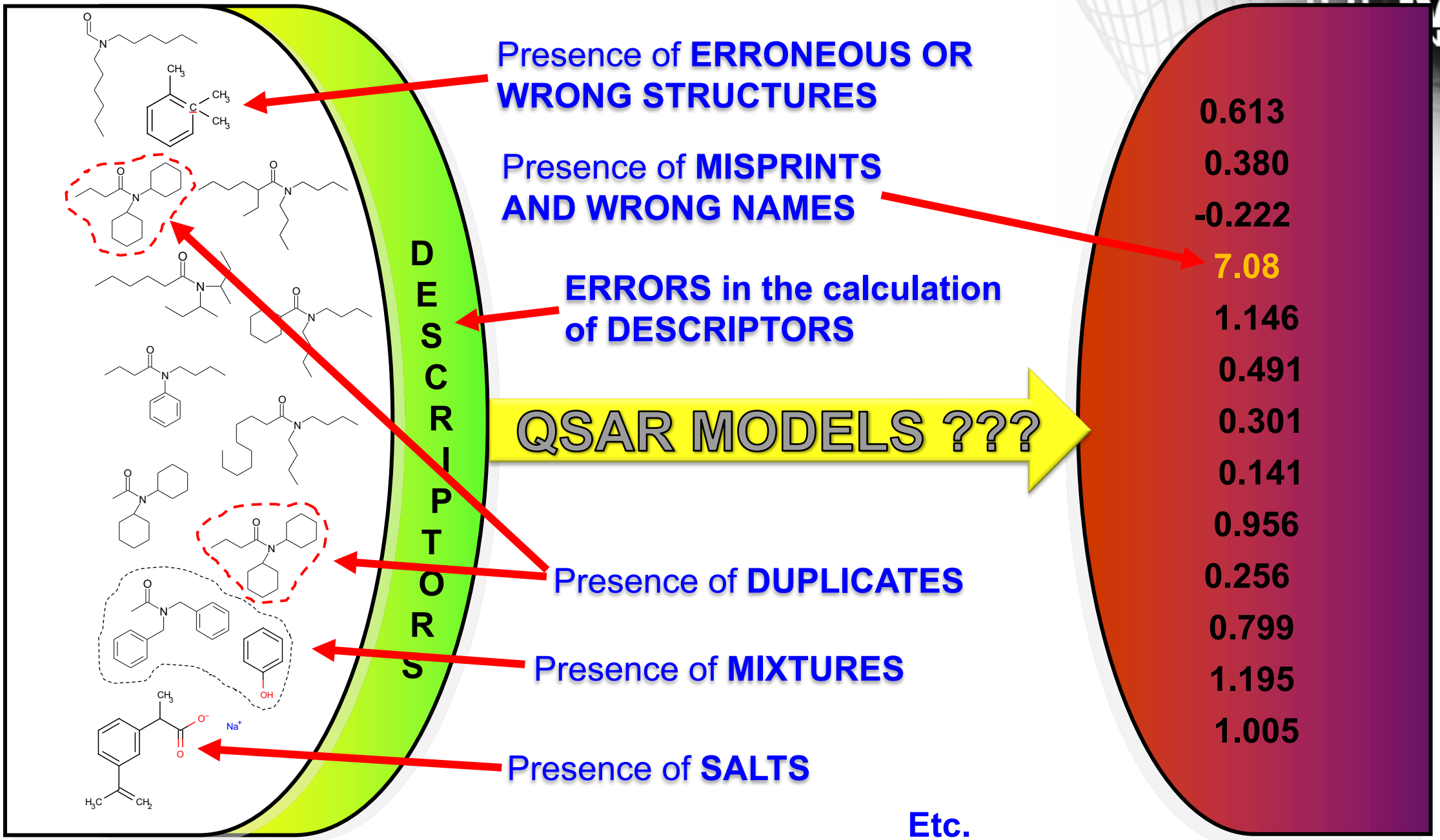
- Cheminformaticians are at the mercy of data providers. Prediction performance of (Q)SAR models could depend strongly on the quality of input data (both structures and activities).

- Both chemical and biological data in a dataset may be inaccurate and in need of thorough curation

- The number of published QSAR models that were poor or not too successful due to data quality issue is unknown but possibly large

- Often considered trivial, the **basic steps to curate a dataset** of compounds are not so obvious especially for beginners.

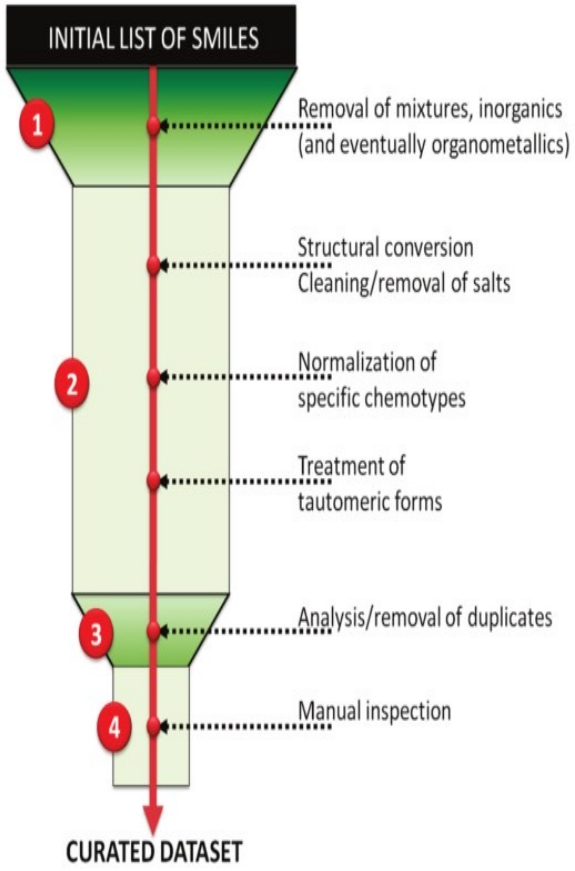
QSAR modeling with non-curated datasets

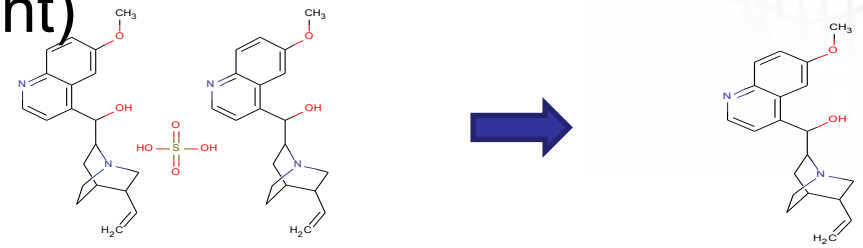


Chemical Structure Curation

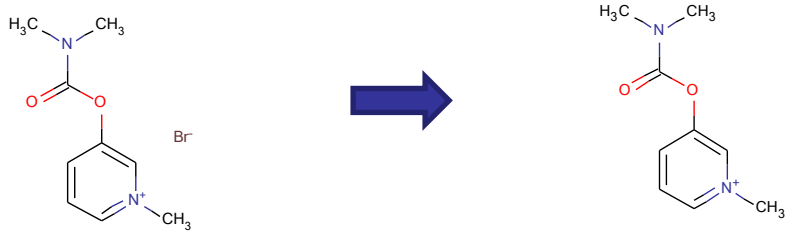
Chemical structures should be cleaned and standardized

(duplicates removed, salts stripped, neutral form, canonical tautomer, etc) to enable rigorous model development)

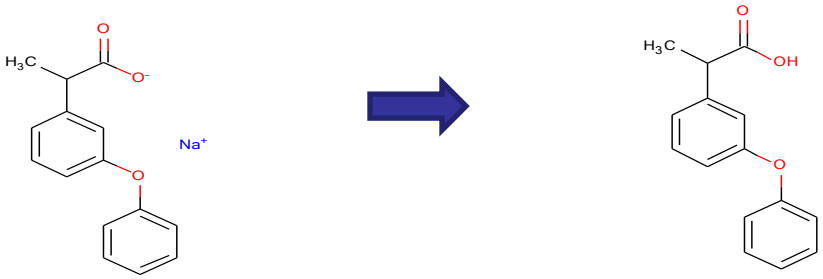




- Quinine sulfate dihydrate



- Pyridostigmine Bromide



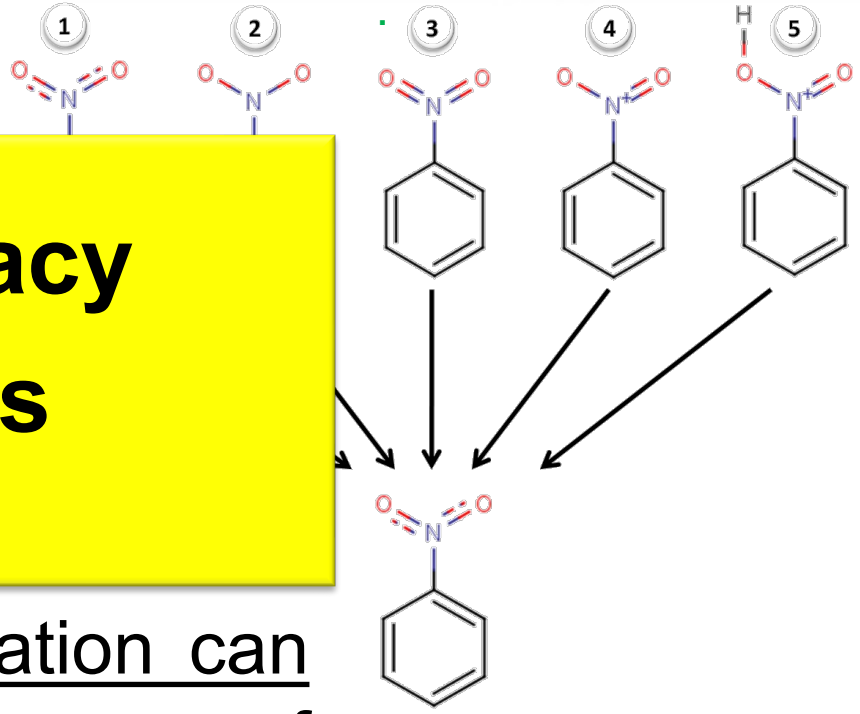
- Fenoprofen Sodium

Muratov, Fourches, Tropsha. Trust but verify. *JC J. Chem. Inf. Model*. **2010**, 50, 1189-1204.

QSAR modeling of nitro-aromatic toxicants

-Case Study 1: 28 compounds tested in rats, log(LD50), mmol/kg.
-Case Study 2: 95 compounds tested against *Tetrahymena pyriformis*, log(IGC50), mmol/ml.

Five different legitimate representations of nitro groups.



Data curation affects the accuracy (up or down!) of QSAR models

Even small differences in structure representation can lead to significant errors in prediction accuracy of models

Artemenko, Muratov et al. SAR QSAR 2011, 22 (5-6), 1-27.

Looking for biological data errors/uncertainties in databases

- What kind of errors do we see?
- When replicate values (of target, ligand, and activity type) appear in the literature, how much do they differ by?
- Does wrong information arise in the laboratory or does it creep in during publication?

Experimental data quality: Comparison of the ToxCAST (Phase I) in vitro Assay Results for Duplicates

Compounds	Total	ACEA	ATG	BSK	Cellumen	NVS	CellzDirect
	500	7	81	87	33	239	48
3-Iodo-2-propynylbutylcarbamate	0.71	0.73	0.18	0.53	0.49	0.89	0.15
Bensulide	0.64	0.09	0.71	0.4	0.69	0.95	0.04
Chlorsulfuron	0.24	N/A	N/A	0.4	N/A	N/A	-0.1
Dibutyl phthalate	0.55	N/A	0.62	0.51	0.7	0.81	-0.1
Diclofop-methyl	0.36	1	0.89	0.15	N/A	-0	-0.1
EPTC	0.13	N/A	N/A	-0.1	N/A	N/A	0.33
Fenoxaprop-ethyl	0.47	N/A	0.56	0.59	0.31	0.35	0.01
Prosulfuron	0.55	N/A	0.68	0.08	N/A	1	0.4

$$^{*}MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

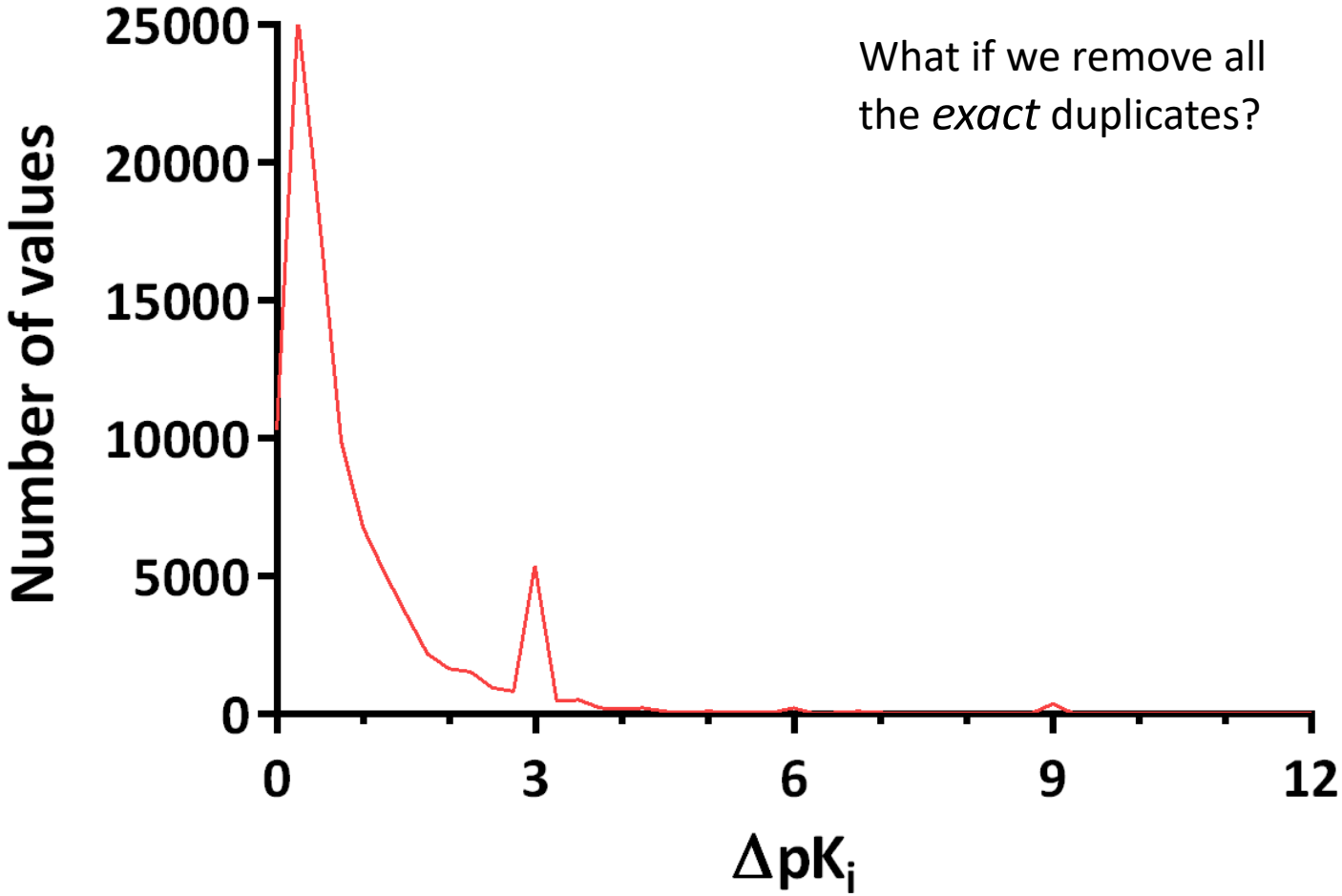
ChEMBL Statistics

- Used ChEMBL 14 – released 18 July 2012
  - 1,384,479 compound records
  - 1,213,242 distinct compounds
  - 644,734 assays
  - 10,129,256 bioactivities
  - 9,003 targets
  - 46,133 documents
- Primarily covers MedChem Literature
- Adds annotations for target data
- Successor to SARLite commercial database

Manual Curation (following several automated steps)

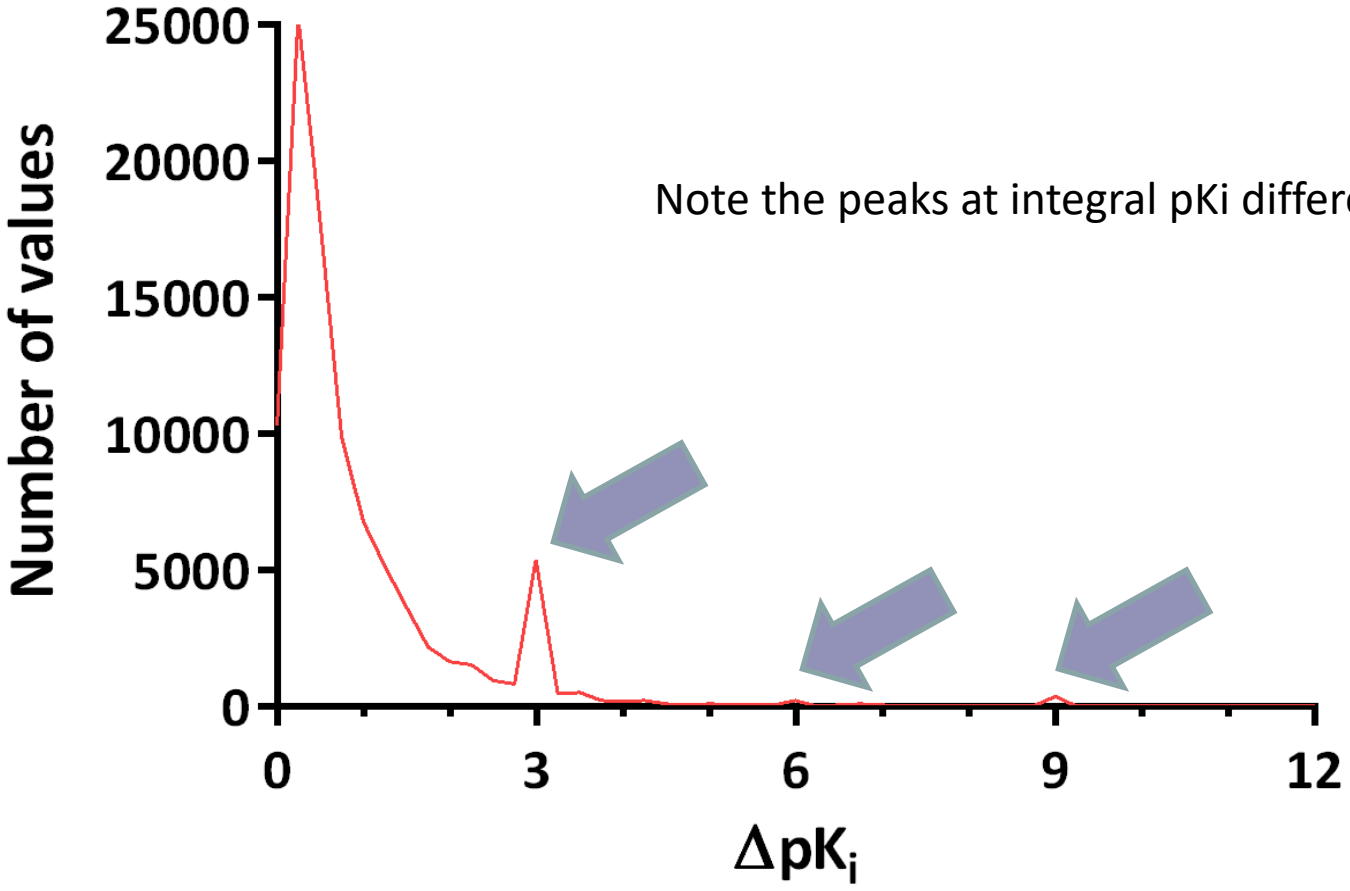
- Input: 190,068 compound-target measures in pairs of papers
  - Used values as published in ChEMBL
  - Converted to standardized pK_i values
  - Semi-automated (based on units and type of value reported)
- 23,956 failed to be automatically converted
  - Mostly Log K_i or –Log K_i values but others
  - Manually examined papers representing ~70% and hand converted affinity value, except when data was being recycled/recited
- Final: 178,317 total replicate pairs of values

Only Replicates > 1% difference



What if we remove all the *exact* duplicates?

A Recurrent Pattern



Note the peaks at integral pKi differences

Non-standard Units Used

J. Med. Chem. **2000**, *43*, 3233−3243

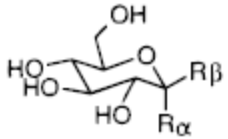
Option

GRid-INdependent Descriptors (GRIND): A Novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors

Manuel Pastor,[†] Gabriele Cruciani,[*,†] Iain McLay,[§] Stephen Pickett,[§] and Sergio Clementi[†]

Laboratory on Chemometrics, Department of Chemistry, University of Perugia, Via Elce di Sotto 10, 06123 Perugia, Italy, and CADD Department, Rhone-Poulenc Rorer, Dagenham, Essex RM10 7XS, U.K.

Table 2. Series of 10 Glucose Analogue Inhibitors of Glycogen Phosphorylase



no.	substituent at C1 position Rα	Rβ	pK_i (mM)
1	OH	H	2.77
2	C(=O)NH_2	H	3.43
3	H	C(=O)NH_2	3.36
4	H	$COOCH_3$	2.55
5	H	CH_2CN	2.05
6	H	NHC(=O)NH_2	3.85
7	C(=O)NH_2	$NHCOOCH_3$	4.80

Non-Ki measures given as Ki

Design, synthesis and structure–activity relationship studies of hexahydropyrazinoquinolines as a novel class of potent and selective dopamine receptor 3 (D_3) ligands

Min Ji[a], Jianyong Chen[a], Ke Ding[a], Xihan Wu[a], Judith Varady[a], Beth Levant[b], Shaomeng Wang[a].

[a] Departments of Internal Medicine and Medicinal Ch
Drive, Ann Arbor, MI 48109-0934, USA

[b] Department of Pharmacology, Toxicology, and The
7417, USA

http://dx.doi.org/10.1016/j.bmcl.2005.01.037, How to

These numbers made it into ChEMBL, too.

Table 1. Binding affinities at the D_1-like, D_2-like and D_3 receptors in binding assays using rat brain

Compounds	K_i ± SEM (nM)			Selectivity	
	D_1-like [^{3}H]SCH 23390	D_2-like [^{3}H]spiperone	D_3 [^{3}H]PD 128907	D_1-like/D_3	D_2-like/D_3
5a	7947 ± 597	3887 ± 664	7487 ± 591	1.1	0.5
5b	8893 ± 568	3643 ± 459	2755 ± 475	3.2	1.3
5c	904 ± 100	243 ± 30	304 ± 53	3.0	0.8
5d	2467 ± 303	852 ± 49	381 ± 59	6.5	2.2
9a	>100,000	>100,000	22,967 ± 6846	>4	>4
9b	356 ± 47	906 ± 190	2523 ± 692	0.1	0.34
9c	258 ± 52	220 ± 21	22 ± 6	12	10
10a	1218 ± 145	1389 ± 111	1650 ± 424	0.7	0.8
10b	152,567 ± 17,284	2443 ± 403	1535 ± 81	10	1
10c	791 ± 187	1568 ± 338	18 ± 2.4	44	87
12a	4602 ± 287	762 ± 51	5.8 ± 1.3	793	131
12b	>250,000	>250,000	244 ± 59	>1000	>1000
12c	5802 ± 422	1125 ± 207	45 ± 7	130	25
12d	6051 ± 570	258 ± 41	2.6 ± 0.4	>2000	99
BP 897	636 ± 103	162 ± 48	1.1 ± 0.2	578	147

Ignorance of Biological Complexity

(8aα,12aα,13aα)-5,8,8a,9,10,11,12,12a,13,13a-Decahydro-3-methoxy-12-(methylsulfonyl)-6*H*-isoquino[2,1-*g*][1,6]naphthyridine, a Potent and Highly Selective α_2-Adrenoceptor Antagonist[1]

J. Med. Chem. **1989**, *32*, 2034–2036

α_2a? α_2b? α_2c?

Table I. Radioligand Binding and Functional Results

compd	ligand binding, pK_i[a]		selectivity[b]
	[^{3}H]prazosin (α_1)	[^{3}H]yohimbine (α_2)	
8a	4.99 ± 0.10	9.18 ± 0.12	15000
8b	5.29 ± 0.10	9.45 ± 0.16	15000
8c	<5	6.32 ± 0.08	>50
idazoxan	6.10 ± 0.08	7.96 ± 0.04	72
yohimbine	6.40 ± 0.03	7.90 ± 0.03	32

Target	Doc_ID	Src_Key	Assay_ID	Activity_ID	Std_Type	Std_Value
α_2a	10218	8b	32635	359172	pK_i	9.45
α_2b	10218	8b	32635	359172	pK_i	9.45
α_2c	10218	8b	32635	359172	pK_i	9.45

No Units at All

Development of High-Affinity 5-HT_3 Receptor Antagonists. 2. Two Novel Tricyclic Benzamides

R. D. Youssefyeh,* H. F. Campbell, J. E. Airey, S. Klein, M. Schnapper, M. Powers, R. Woodward, W. Rodriguez, S. Golec, W. Studt, S. A. Dodson, L. R. Fitzpatrick, C. E. Pendley, and G. E. Martin

Rhône-Poulenc Rorer Central Research, 640 Allendale Road, King of Prussia, Pennsylvania 19406. *Received August 23, 1991*

Table II. Antagonism of [^{3}H]GR 65630 Binding by Various Agents

compd	n^a	$K_i \pm$ SE	compd	n^a	$K_i \pm$ SE
8	1	1.07 ± 0.57	**24**	1	>100
9	2	0.74 ± 0.14	**25**	1	>100
10	7	0.17 ± 0.02	**26**	1	>100
11	3	8.77 ± 1.82	**27**	1	29.6 ± 5.7
12	3	2.05 ± 0.12	**28**	1	>100
13	2	2.85 ± 1.16	BRL 43694	3	1.72 ± 0.03
18	1	0.30 ± 0.14	GR 38032F	3	6.16 ± 2.1
19	1	3.42 ± 0.84	ICS 205-930	5	2.1 ± 0.50
20	1	1.96 ± 0.55	MDI 72222	3	21.12 ± 8.6
21	2	0.69 ± 0.23	zacopride	3	1.51 ± 0.36

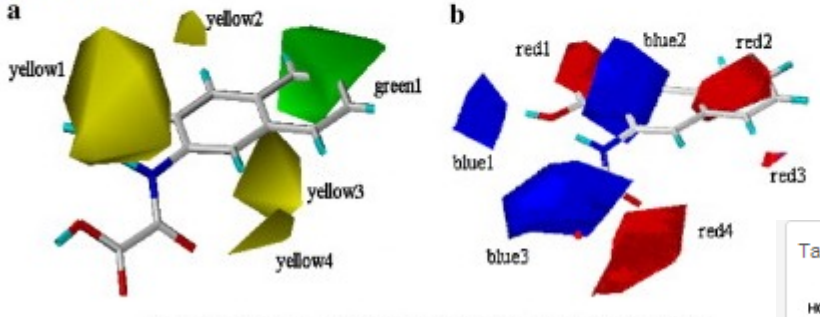
^{a} n = number of experiments. On each experiment compounds were tested in six-point competition experiments with triplicate replication.

No Citation For Data Sources

Molecular docking and 3D-QSAR on 2-(oxalylamino) benzoic acid and its analogues as protein tyrosine phosphatase 1B inhibitors
Pages 5521-5525
Mei Zhou, Mingjuan Ji

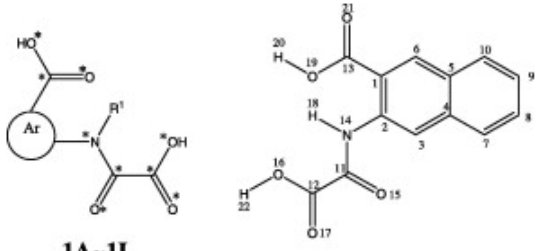
Show preview | PDF (231 K) | Related articles | Related reference work articles

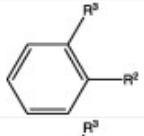
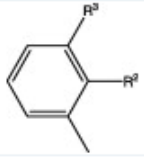
Graphical abstract
The figure showed the inhibitor modification information derived from CoMFA model. Increasing bulk inside green regions and removing bulk from yellow regions favor the inhibitory activity; increasing negative charge in red regions and increasing positive charge in blue regions favor the inhibitory activity.



The contour plots of CoMFA steric fields (a) and electrostatic fields (b)

Table 1. Structure and activity data of inhibitors for CoMFA and FlexX

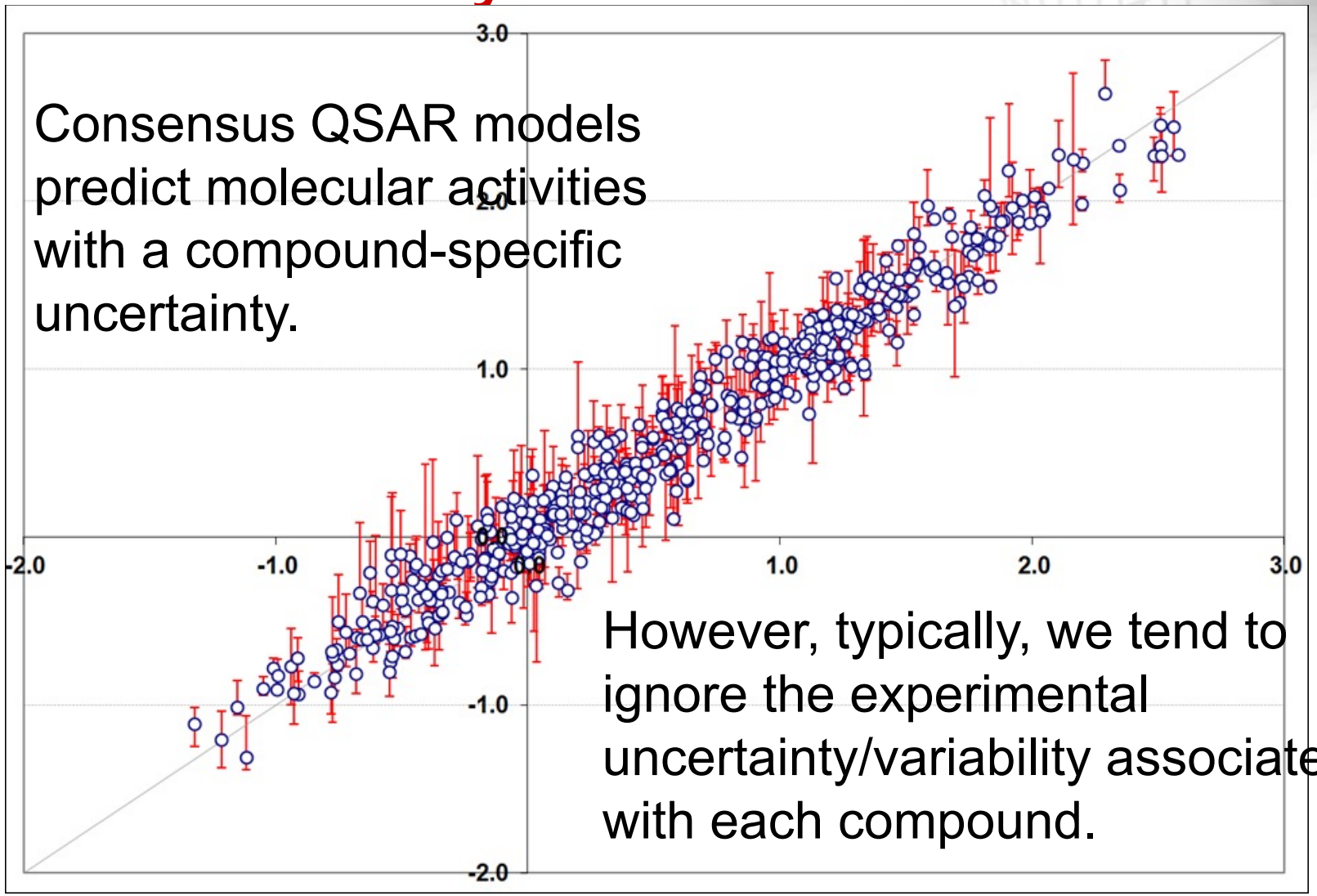


Compound	Ar[a]	R^1	pK_i^b(obsd)	pK_i (calcd)	Residue	Total score (kJ mol^{-1})
1A		H	4.638	3.720	0.919	−154.03
1B		H	2.959	3.045	−0.086	−126.82

Summary of published data quality analysis

- A lot of the replicates in the literature aren't actually independent determinations
- Many errors come from careless specification or interconversion of units
- 91% of the data are single reported measurements
- Modeling studies often are not explicitly identified as such
- ChEMBL 15 and going forward have started to address these issues
- This observations suggest new challenges to employ cheminformatics approaches for biological data curation

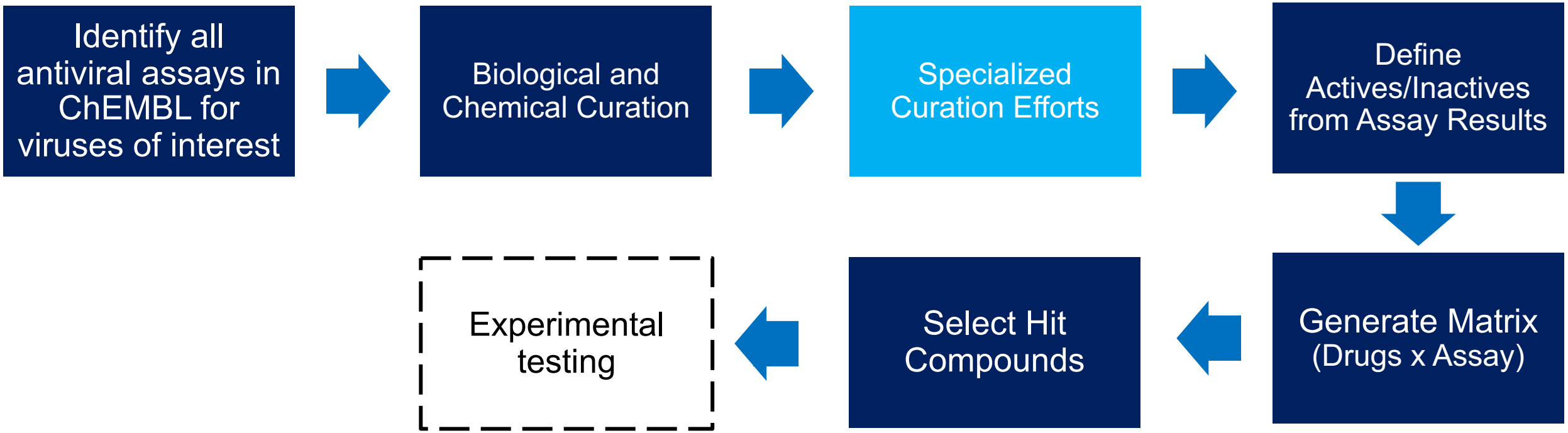
ChEMBL Statistics: experimental uncertainty



Consensus QSAR models predict molecular activities with a compound-specific uncertainty.

However, typically, we tend to ignore the experimental uncertainty/variability associated with each compound.

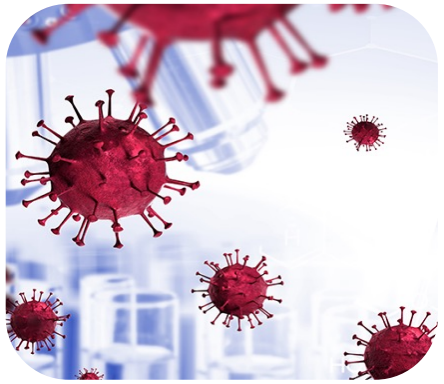
Recent curation effort: creation of a derivative database of antiviral compounds found in ChEMBL



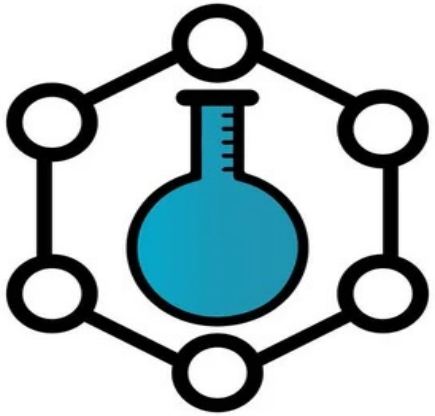
Martin, H. et al, Antiviral Res., 2023 Sep;217:105620.

Seems easy! Just look it up in the ChEMBL database… right?

BUT: Grave issues with ChEMBL's antiviral assay ontology and annotation…



Assay Type



Assay Conditions

Total Time Spent Fixing These Issues: ~75 hours

Assay Ontology Issues: Assay Descriptions

Inclusions: virus, cell, assay, time, concentration, assessment

Worst → Best

Antiviral activity against SARS-CoV-2

Antiviral activity determined as inhibition of SARS-CoV-2 in HeLa cells

MOST COMMON

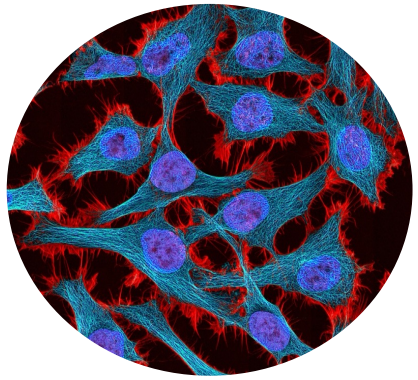
Antiviral activity determined as inhibition of SARS-CoV-2 induced cytotoxicity of VERO-6 cells at 10 uM after 48 hours exposure to 0.01 MOI SARS CoV-2 virus by high content imaging

Virus Info Only

Virus + 1-2 Inclusions

All Inclusions

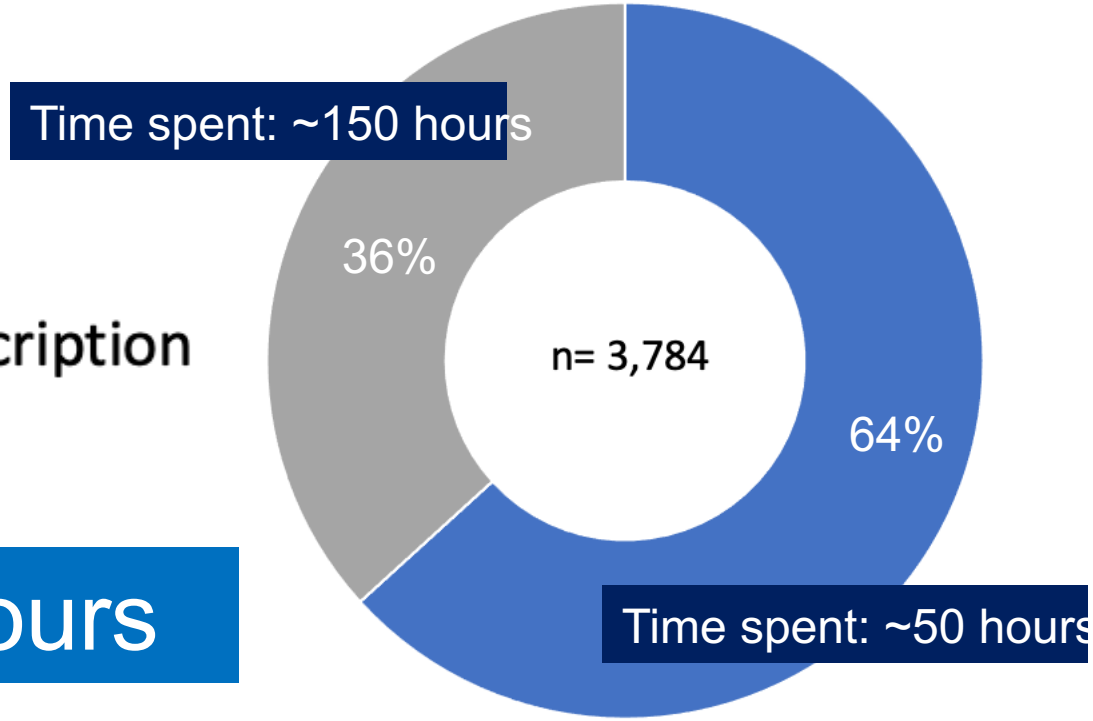
Heavy curation efforts: for instance, missing cell types in phenotypic assays



Cell Type

14% of all phenotypic assay results

were missing the cell-type from the designated field



- Found in Assay Description
- Completely Missing

Total time spent: ~200 Hours

BAO Mislabeling Impacts Data Accessibility

Using the “BAO Assay Type” as a filter to search ChEMBL for cell-based assay’s for my viruses of interest **would have cost 99.44% of all collected data.** It was effectively HIDDEN!

Total time spent:
~25 Hours

Summary of antiviral compound activity in curated subset from ChEMBL

32,515 compound entries x 13 viruses

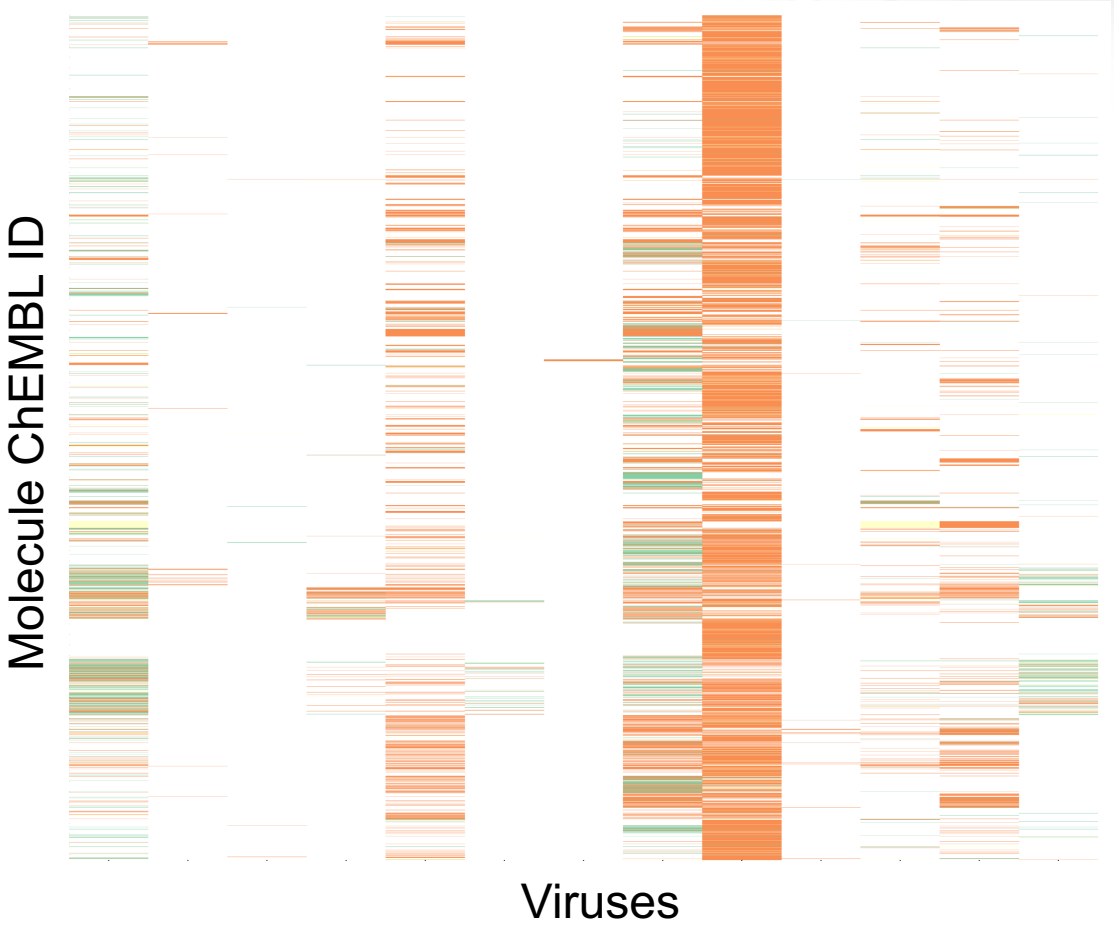
Thresholds
% inhibition > 50
EC50 ≤ 10 μM
IC50 ≤ 10 μM

Active

Inactive

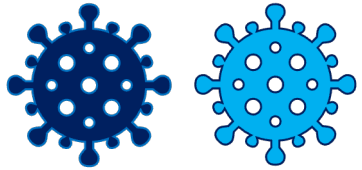
Inconclusive

Not tested



New Testing Recommendations

Criteria



Active in 1+ phenotypic assay(s) in 2+ different viruses

Hypothesis

Broad-Spectrum for Viral Family

Compound Profile Example

Compound ID	Phenotypic Activity	Phenotypic Inactivity	Untested Phenotypic	Untested Target-Based
Compound X	Dengue 1; Zika	None	Yellow Fever; West Nile; Dengue 2-4	All

Testing Recommendations

1. Retest nominated compounds against Dengue 1 and Zika to ensure assay compatibility
2. Test against Dengue 2-4, West Nile, and Yellow Fever due to high conservation amongst flavivirus proteins

Flavivirus Screening Results

MML UNC.EDU

- 73 compounds tested at DENV 2&4 (some with reported DENV activity, some with activity at other flaviviruses)
- **Total of 43 unique compounds (+4-5 controls) had significant activity <50% RLU):**

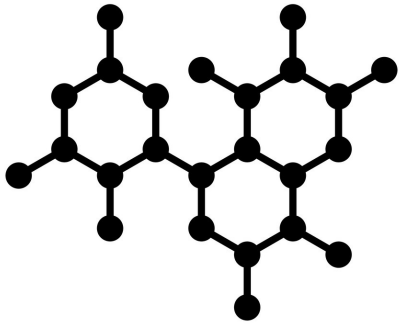
Virus and Assay Concentration	# of compounds active	% of compounds tested
DENV2nLuc (% RLU) 1uM	13	17.8%
DENV4nLuc (% RLU) 1uM	10	13.6%
DENV2nLuc (% RLU) 10uM	46	63.0%
DENV4nLuc (% RLU) 10uM	40	54.7%

Finally! Small Molecule Antiviral Compound Collection (SMACC)*

Drug-Assay Pairs

32,515

Compounds



12,221

Assays

1,119

The picture can't be displayed.

*https://smacc.mml.unc.edu.

*Martin, H. et al, Antiviral Res., 2023 Sep;217:105620.

Analysis of one publication: CYP data

Predictive Models for Cytochrome P450 Isozymes Based on Quantitative High Throughput Screening Data

Hongmao Sun,*,† Henrike Veith,† Menghang Xia,† Christopher P. Austin,† and Ruili Huang†

†National Institutes of Health (NIH) Chemical Genomics Center, NIH,

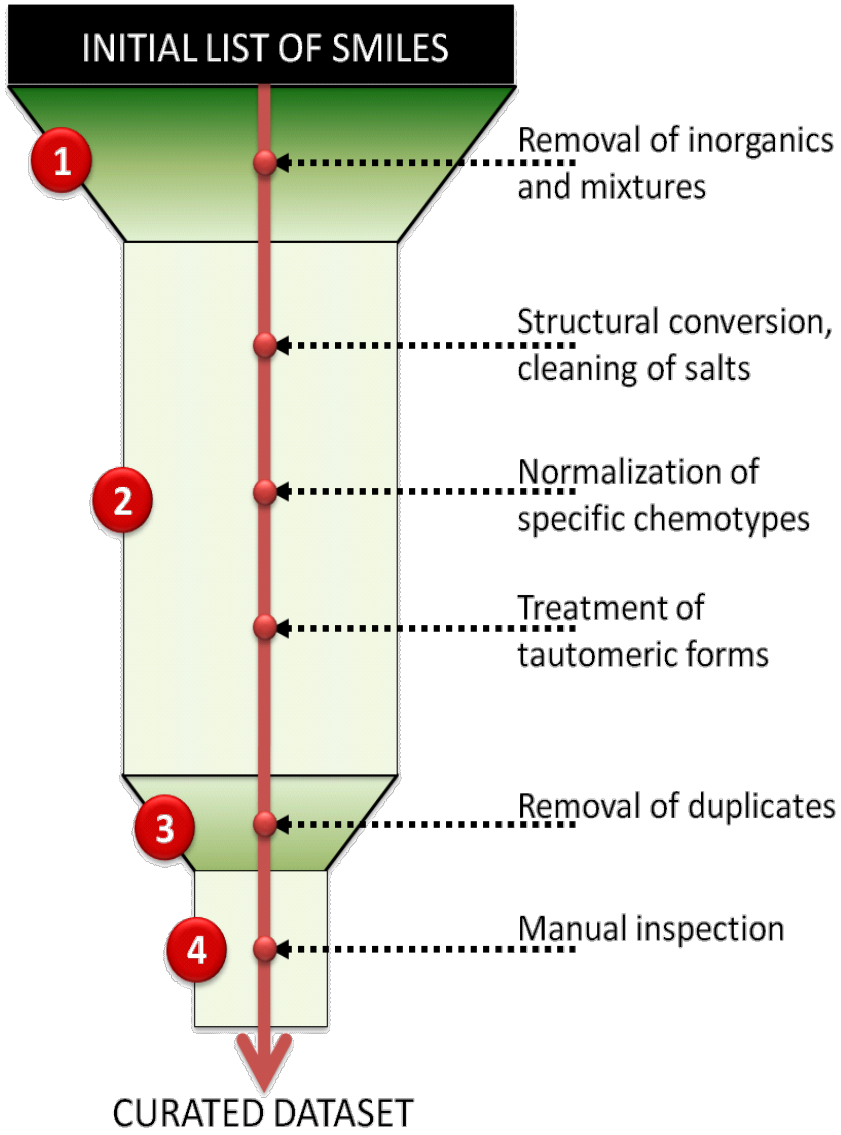
ABSTRACT: The human cytochrome P450 (CYP450) isozymes are the most important enzymes in the body to metabolize many endogenous and exogenous substances including environmental toxins and therapeutic drugs. Any unnecessary interactions between a small molecule and CYP450 isozymes may raise a potential to disarm the integrity of the protection. Accurately predicting the potential interactions between a small molecule and CYP450 isozymes is highly desirable for assessing the metabolic stability and toxicity of the molecule. The National Institutes of Health Chemical Genomics Center (NCGC) has screened a collection of over 17,000 compounds against the five major isozymes of CYP450 (1A2, 2C9, 2C19, 2D6, and 3A4) in a quantitative high throughput screening (qHTS) format. In this study, we developed support vector classification (SVC) models for these five isozymes using a set of customized generic atom types. The CYP450 data sets were randomly split into equal-sized training and test sets. The optimized SVC models exhibited high predictive power against the test sets for all five CYP450 isozymes with accuracies of 0.93, 0.89, 0.89, 0.85, and 0.87 for 1A2, 2C9, 2C19, 2D6, and 3A4, respectively, as measured by the area under the receiver operating characteristic (ROC) curves. The important atom types and features extracted from the five models are consistent with the structural preferences for different CYP450 substrates reported in the literature. We also identified novel features with significant discerning power to separate CYP450 actives from inactives. These models can be useful in prioritizing compounds in a drug discovery pipeline or recognizing the toxic potential of environmental chemicals.

CONCLUSION

SVM classification models have been built for the five most important isoforms of CYP450 (1A2, 2C9, 2C19, 2D6, and 3A4) based on a large qHTS data set with over 6000 compounds available for both model training and testing. The five CV optimized SVC models built by using the atom typing molecular descriptors exhibited consistently high predictive power when applied to the equally populated test sets with accuracies between 0.85 and 0.93, as measured by the AUC of ROC plots. The results indicated that the atom typing descriptors generated from a large, high quality data set were capable of feeding information rich learning materials to the SVM learner. Useful information of structural features was derived from feature importance analysis for each isozyme of CYP450. The privileged structural features that could result in inhibitory and stimulatory activity against different CYP450 isozymes can serve as valuable guidelines in the drug discovery process.

J. Chem. Inf. Model. 2011, 51, 2474–2481

Dataset Curation summary



	17143 compounds
Removal of inorganics and mixtures	17121 compounds
Structural conversion, cleaning of salts	17121 compounds
Normalization of specific chemotypes	17121 compounds
Treatment of tautomeric forms	17121 compounds
Removal of duplicates	16142 compounds
Manual inspection	**16142** compounds

Fourches D, et al. J Chem Inf Model. 2010 50(7):1189-204.

NCGC dataset analysis of duplicates

- Out of 1280 duplicate couples :
  - 406 had no discrepancies-no values or no values for comparison
  - 874 had biological profile differences

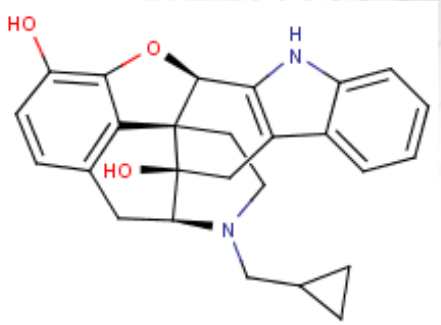
- A total of 1535 discrepancies were found in the 874 couples of duplicates:

	CYP2C9	CYP1A2	CYP3A4	CYP2D6	CYP2C19
# of discrepancies	154	363	426	422	170

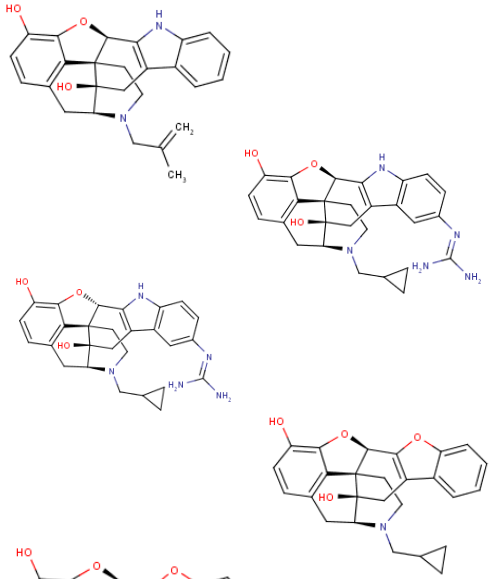
Neighborhood Analysis for Duplicates

17,000 compounds screened against five major CYP450 isozymes.
1,280 pairs of duplicates couples were found (874 had different bioprofiles)

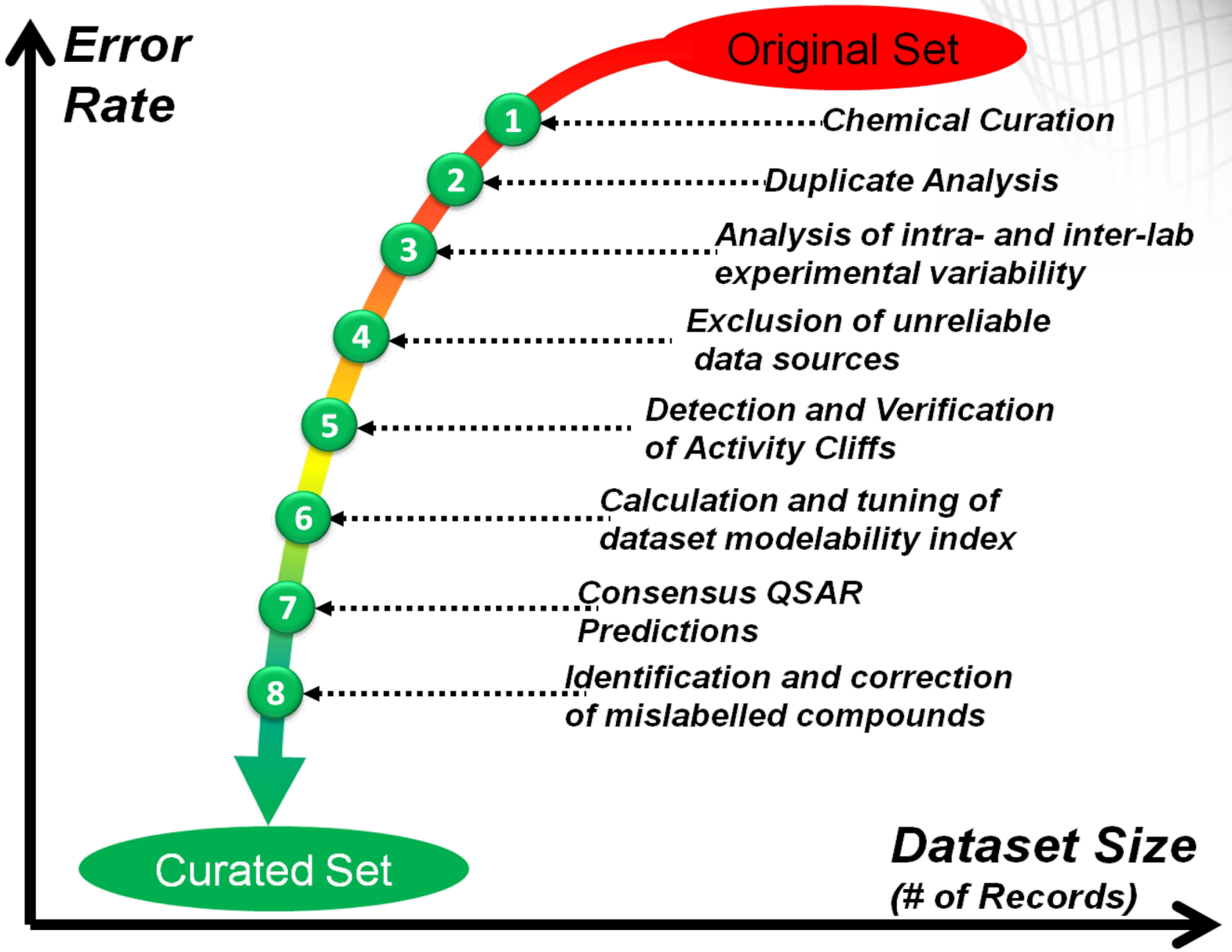
Tocris-0740	SID	Supplier	2C9	1A2	3A4	2D6	2C19
CID_6603937	11113673	Tocris	-4.6	-4.4	**-4.6**	-6.2	-4.5
CID_6603937	11111504	Sigma Aldrich	-4.4		**-8**	-5.6	-5



5 Nearest neighbors	Tanimoto Similarity	SID	Supplier	2C9	1A2	3A4	2D6	2C19
6604862	**0.98**	11114071	Tocris			**-4.5**		-5.5
6604106	**0.98**	11112029	Sigma Aldrich			**-5.1**		
6604846	**0.98**	11114012	Tocris					
6604136	**0.95**	11112054	Sigma Aldrich			**-4.8**	-5.9	
6604137	**0.95**	11113764	Tocris		-4.4	**-4.7**	-4.5	



Biological data curation workflow



Fourches et al., Curation of Chemogenomics Data. Nature Chem. Bio., 2015, in press.

Notes on the importance of data curation

- The curation of chemical data is critical prior to any cheminformatics analysis and modeling. Difficult cases require human interventions and cannot be fully automated.

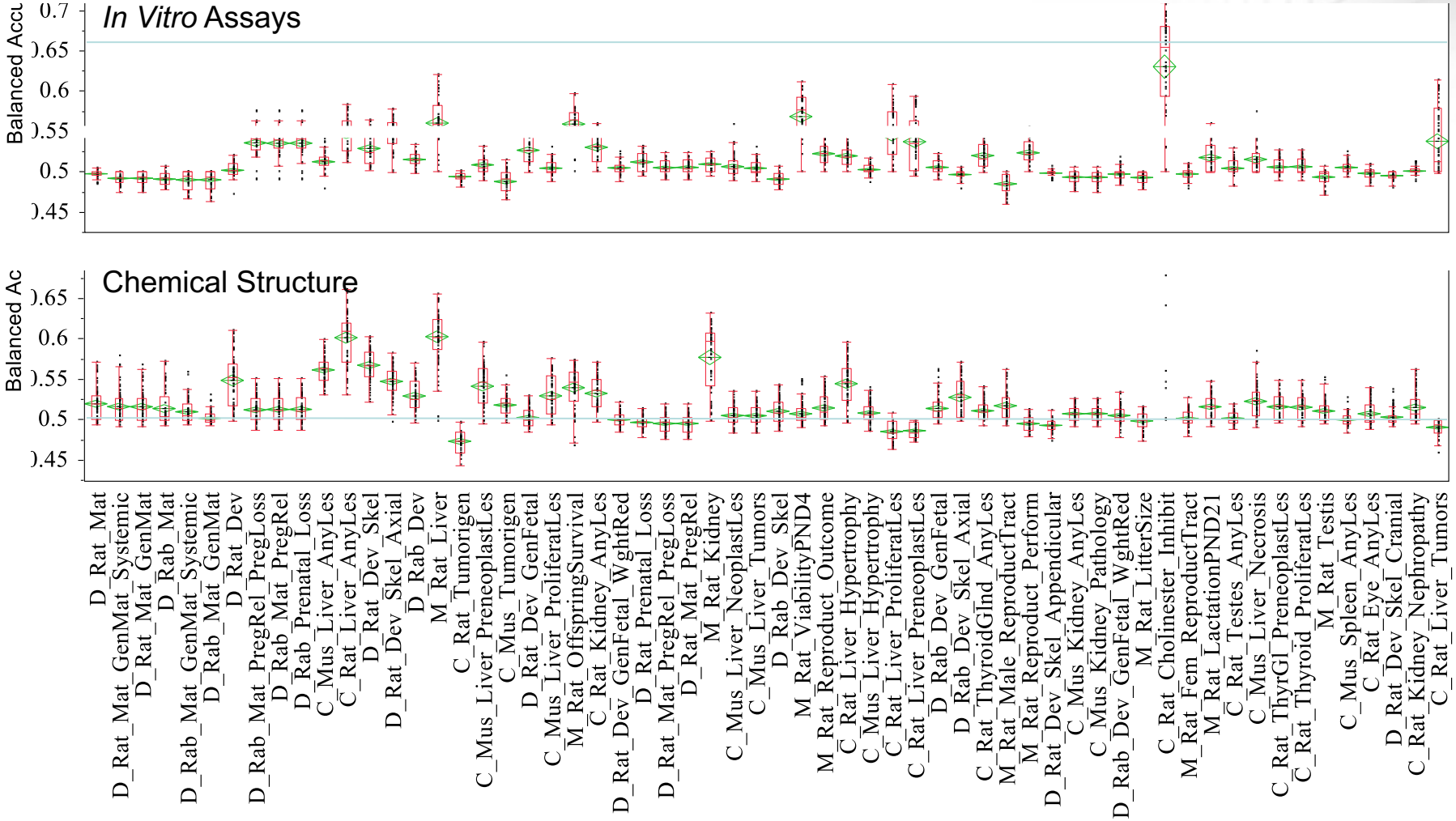
- Prediction outliers may be due to structural outliers, real activity cliffs or mislabeled compounds. Many of them can still be detected and removed prior to modeling studies boosting the reliability of QSAR model.

- Rigorously developed QSAR models can be used to correct erroneous biological data associated with certain compounds.

Free and open-source QSAR-ready workflow for automated standardization of chemical structures in support of QSAR modeling. Mansouri et al, J Cheminform . 2024 Feb 20;16(1):19. doi: 10.1186/s13321-024-00814-3.

Dataset Modelability: does it make sense to model any SAR data?

Example: Poor *structure – in vivo* or *in vitro-in vivo* correlations for Toxcast data*



Toxicol Sci. 2012 Aug;128(2):398-417.

The Concept of Modelability

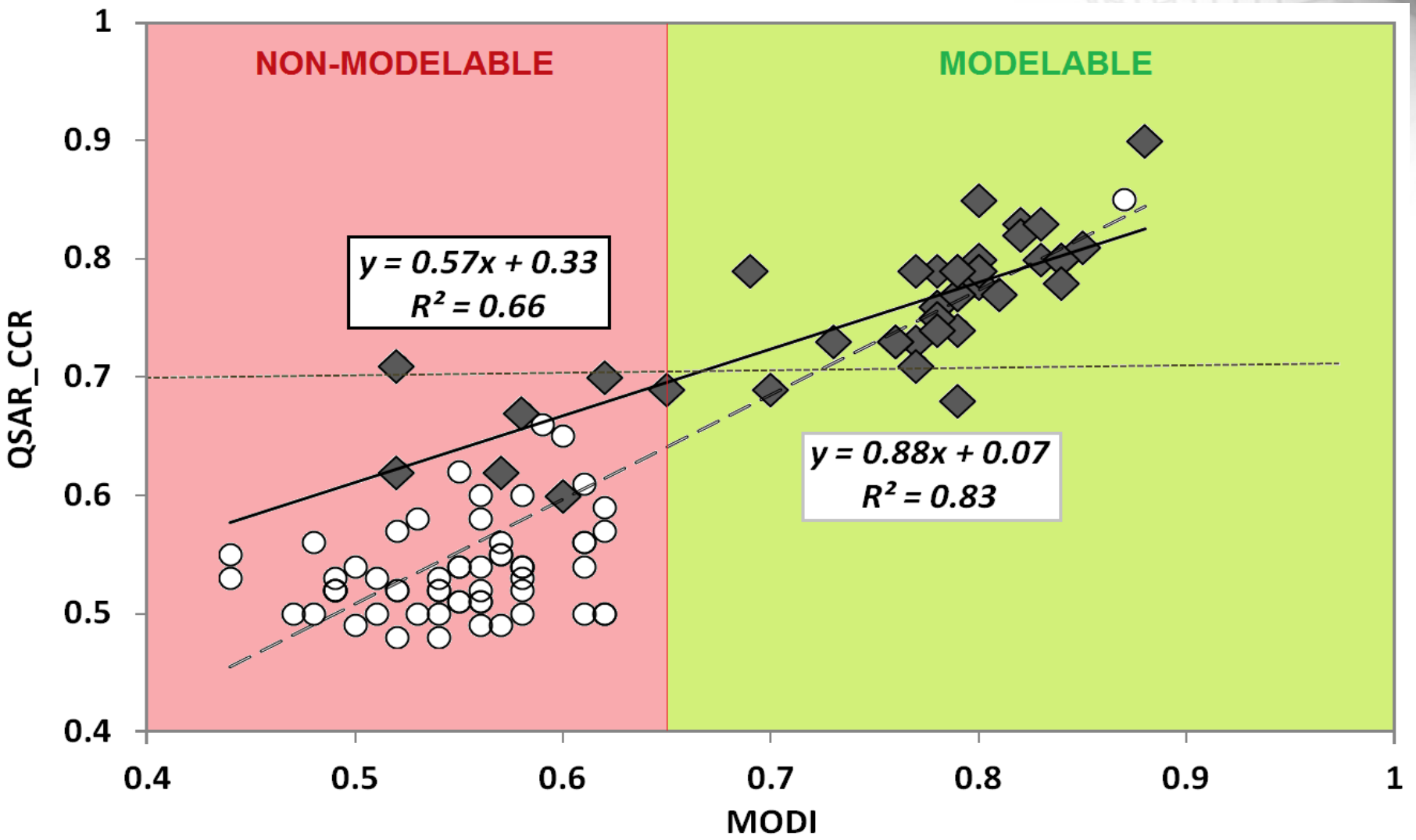
- We often fail to build a predictive QSAR model. However, it may be possible to evaluate ***modelability*** of the dataset prior to QSAR study.

$$\mathrm{MODI} = \frac{1}{K}\sum_{i=1}^{K}\frac{N_i^{\mathrm{same}}}{N_i^{\mathrm{total}}}$$

where K is the number of classes ($K = 2$ for binary data sets), N_i^{same} is the number of compounds of i-th activity class that have their first nearest neighbors belonging to the same activity class i; N_i^{total} is the total number of compounds belonging to the class i.

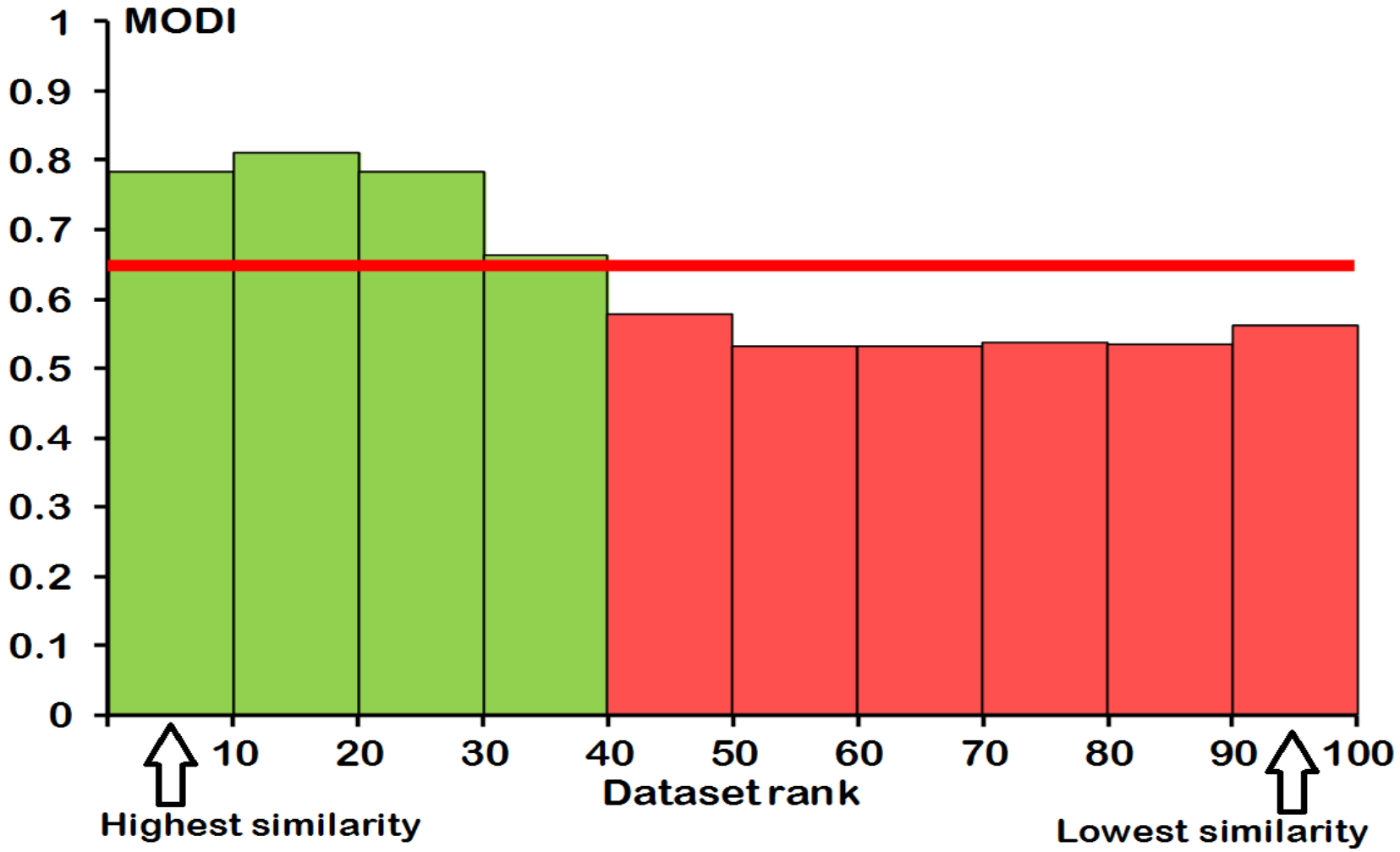
*Golbraikh et al. Data Set Modelability by QSAR. *J Chem Inf Model*. 2014, 54, 1-4

Prediction of dataset modelability



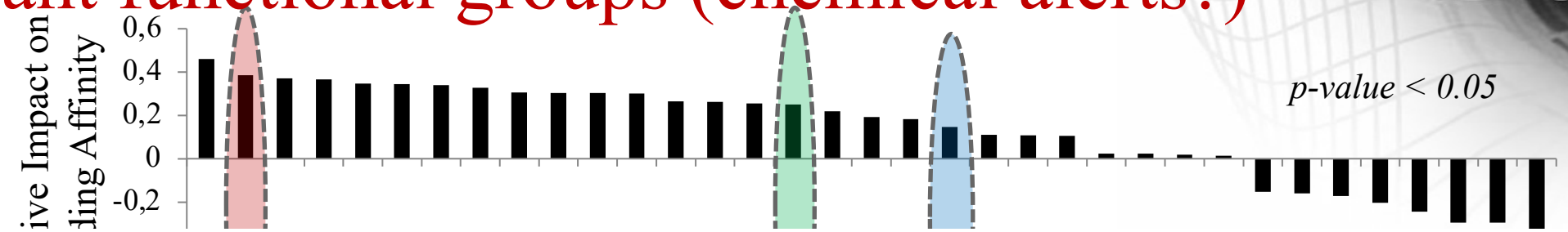
*Golbraikh et al. Data Set Modelability by QSAR. *J Chem Inf Model*. 2014, 54, 1-4

Modelability and Structural Dissimilarity

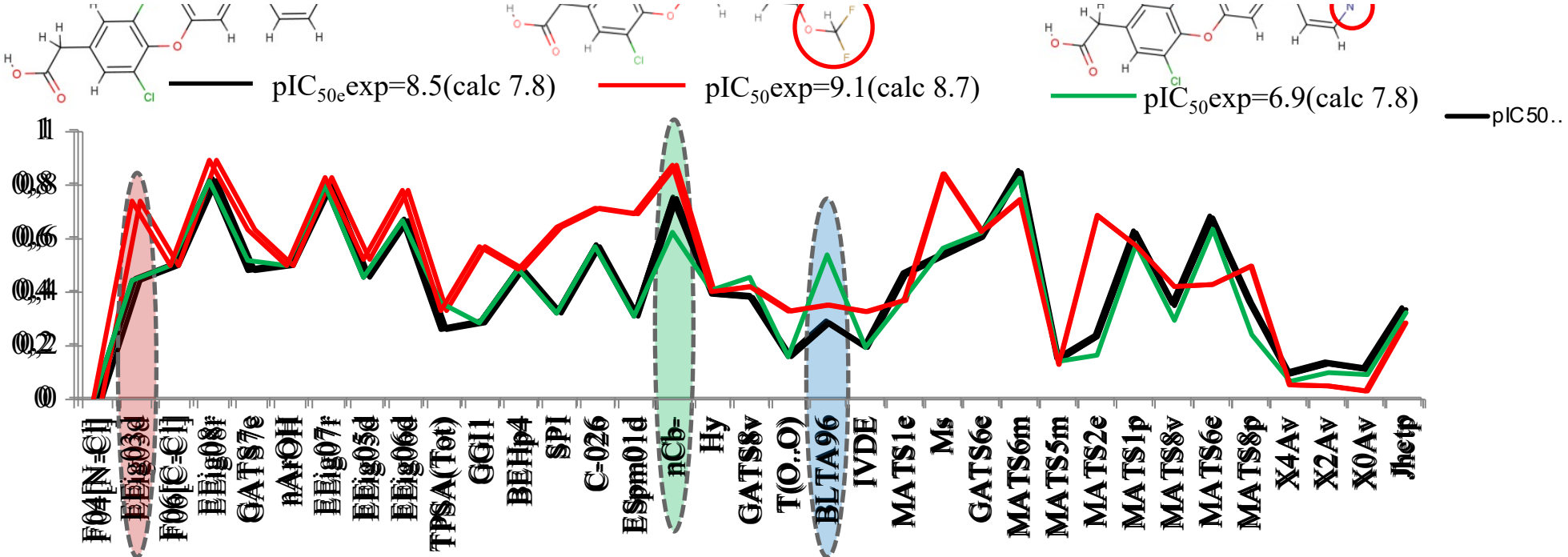


*Golbraikh et al. Data Set Modelability by QSAR. *J Chem Inf Model*. 2014, 54, 1-4

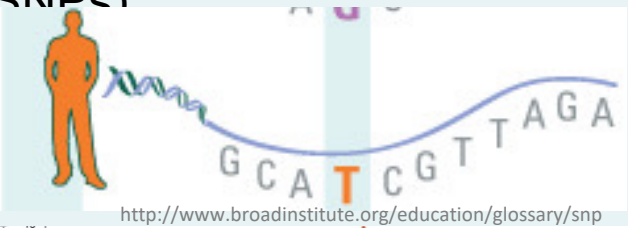
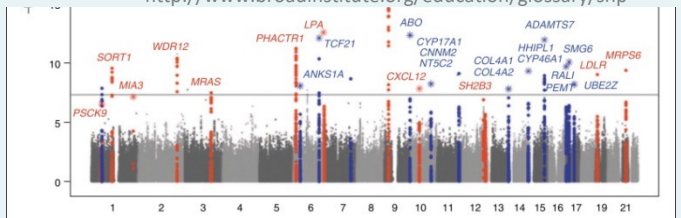
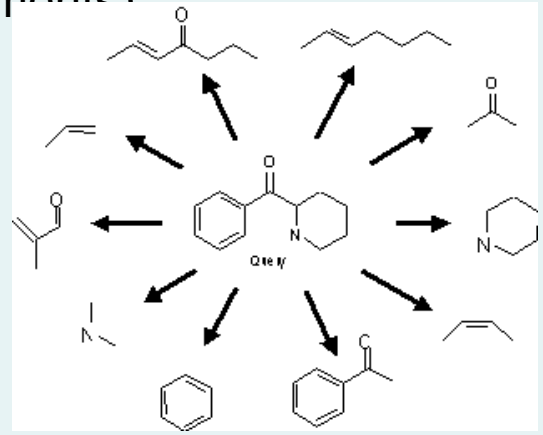
Can QSAR models be interpreted in terms of significant functional groups (chemical alerts?)



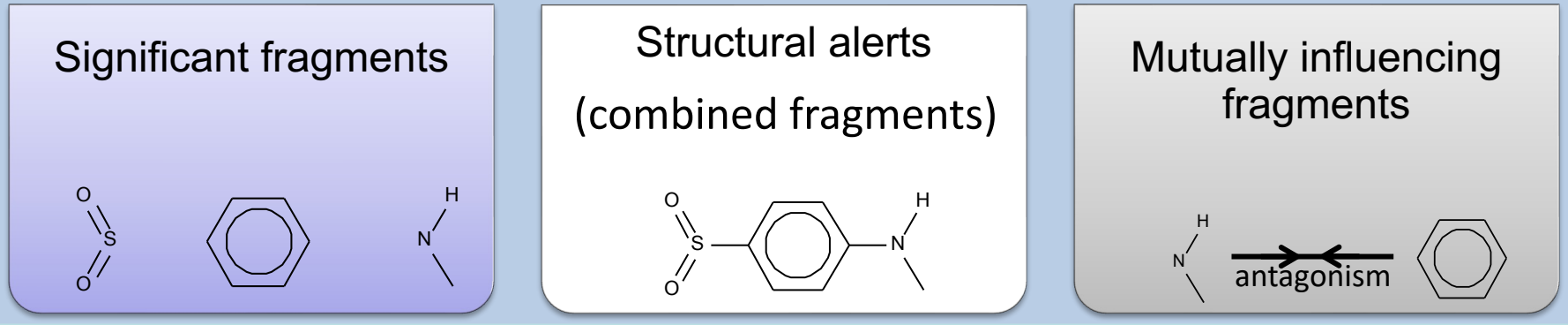
Important observation: chemical features never act in isolation from the rest of the structure! Explanation of multivariate models by one or few descriptors is typically non-sensible



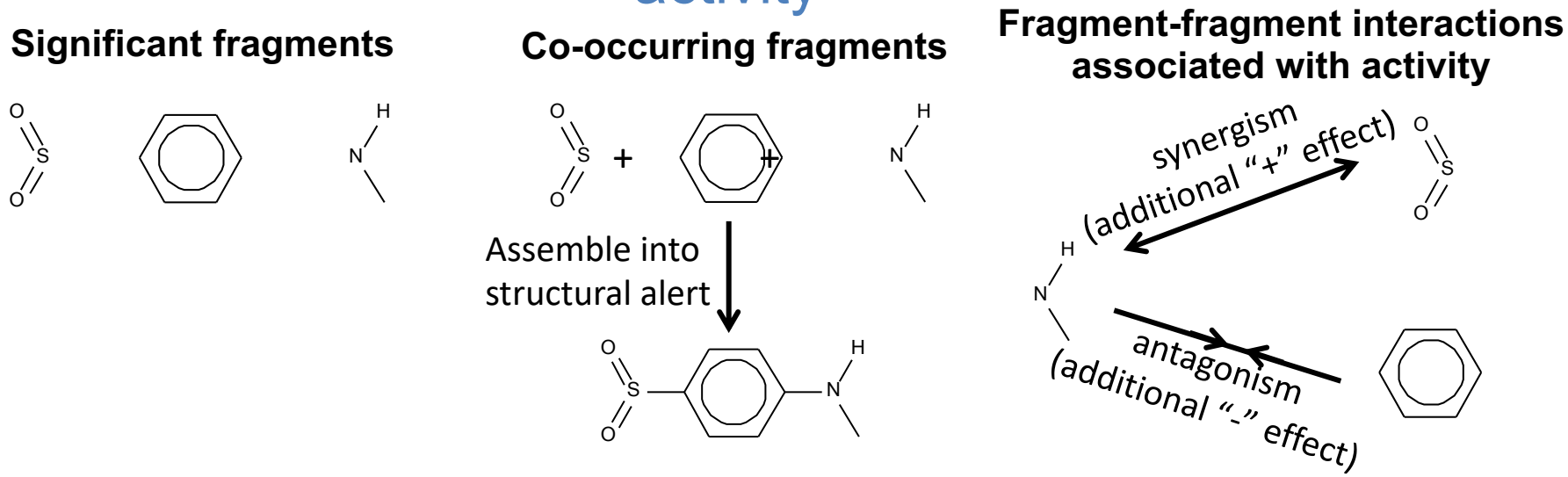
Model interpretation based on Chemistry-Wide Association Studies (CWAS)

	GWAS	(Q)SAR
Samples	Patients	Compounds
Response	Phenotype (disease/no disease)	Activity (active/inactive)
Features	Single Nucleotide Polymorphisms (SNPs)  http://www.broadinstitute.org/education/glossary/snp  Lusis A, Genetics of atherosclerosis, Trends in Genetics (2012) 28(6):267-275	Chemical descriptors (e.g. fragments)  http://www.aldrichmarketselect.com/support/similarityOverview.asp
Objectives	Identify SNPs/loci associated with phenotype Predict phenotype from SNPs	Identify substructure associated with activity Predict activity from structure

CWAS: develop and employ QSAR models using GWAS framework



CWAS: study how chemical structures are associated with activity



Modeling and identifying important fragments

Ames data set
5,439 compounds
2,121 mutagenic
3,318 non-mutagenic

967 fragments → 76 fragments

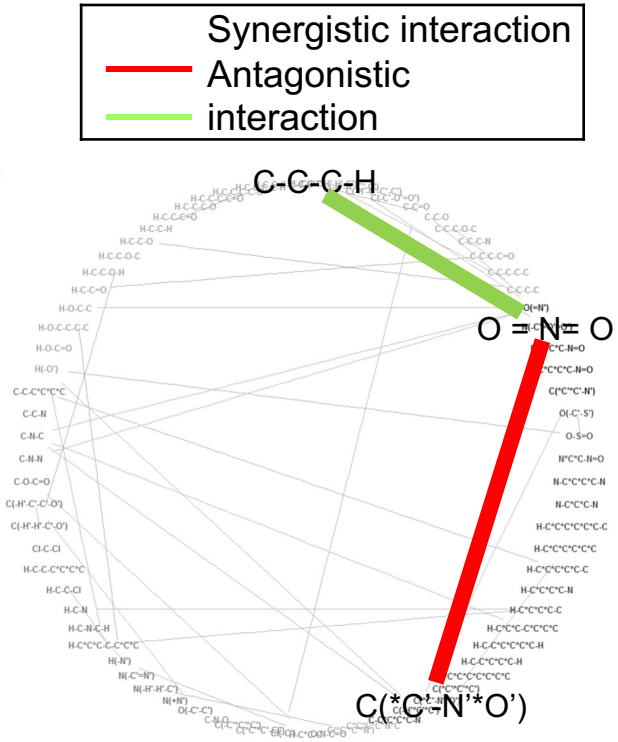
Data preprocessing → Build random forests (RF) → Rank fragments by RF importance → Rebuild RF models with top fragments

Chemical curation
Remove invariant, highly correlated fragments

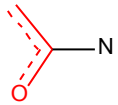
	Full model (967 fragments)		Reduced model (76 fragments)
Specificity	0.92 ±0.009	Slightly improved	0.92 ±0.009
Sensitivity	0.78 ±0.005		0.81 ±0.005
Balanced Accuracy	0.85 ±0.005		0.87 ±0.005
AUC	0.91 ±0.004		0.94 ±0.003

Results from 5-fold external cross validation

Nitro's mutagenic effect is:
increased by furan (synergism)
decreased by primary alkanes (antagonism)

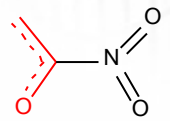


Synergistic influence

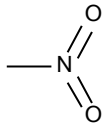


94% mutagenic
79:5

+



100% mutagenic
79:0



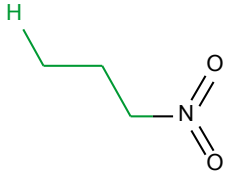
84% mutagenic ("penetrance")
620:118

+

C – C – C – H

29% mutagenic
785:1884

Antagonistic influence

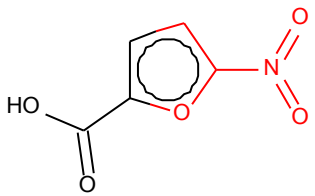


69% mutagenic
100:46

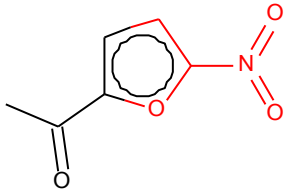
Number of mutagenic compounds : Number of non-mutagenic compounds

Nitro compounds are active when paired with aromatic rings and inactive when paired with primary alkanes

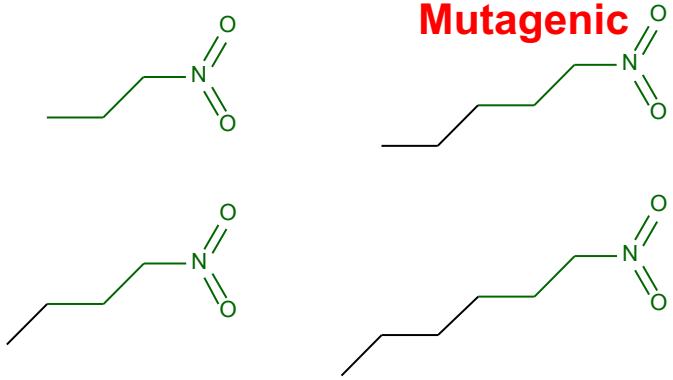
Examples



645-12-5
5-nitro-2-furanoate
Mutagenic

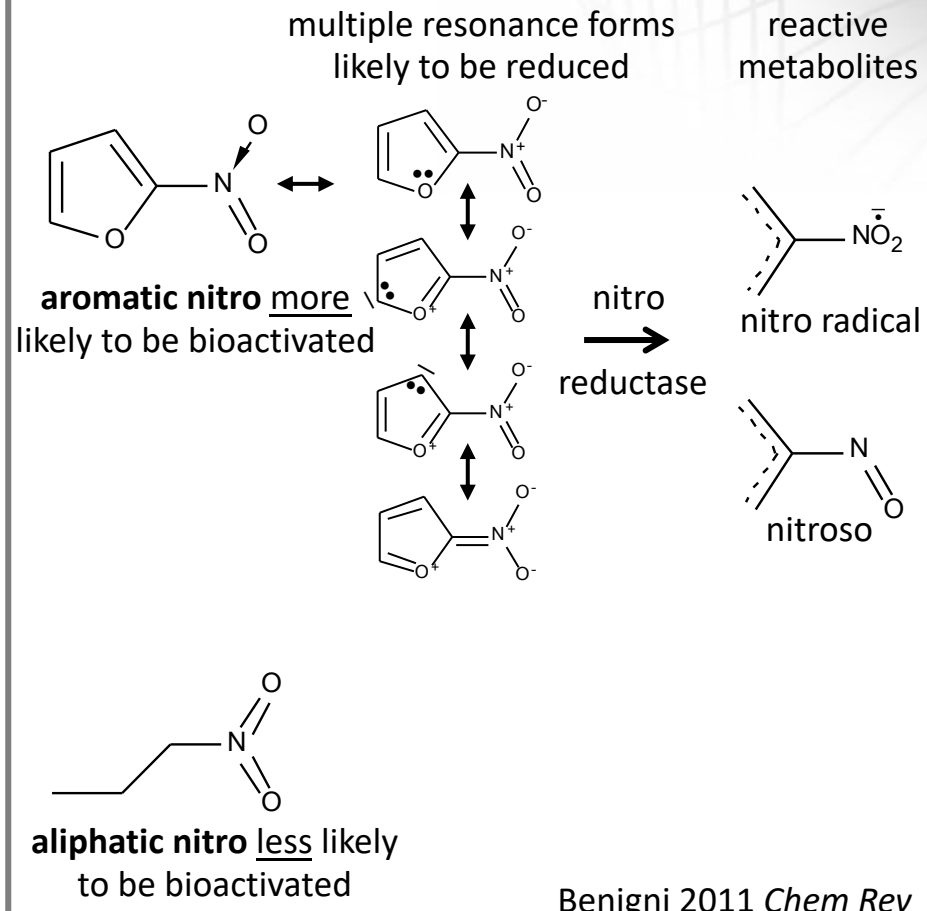


5275-69-4
2-acetyl-5-nitrofuran
Mutagenic



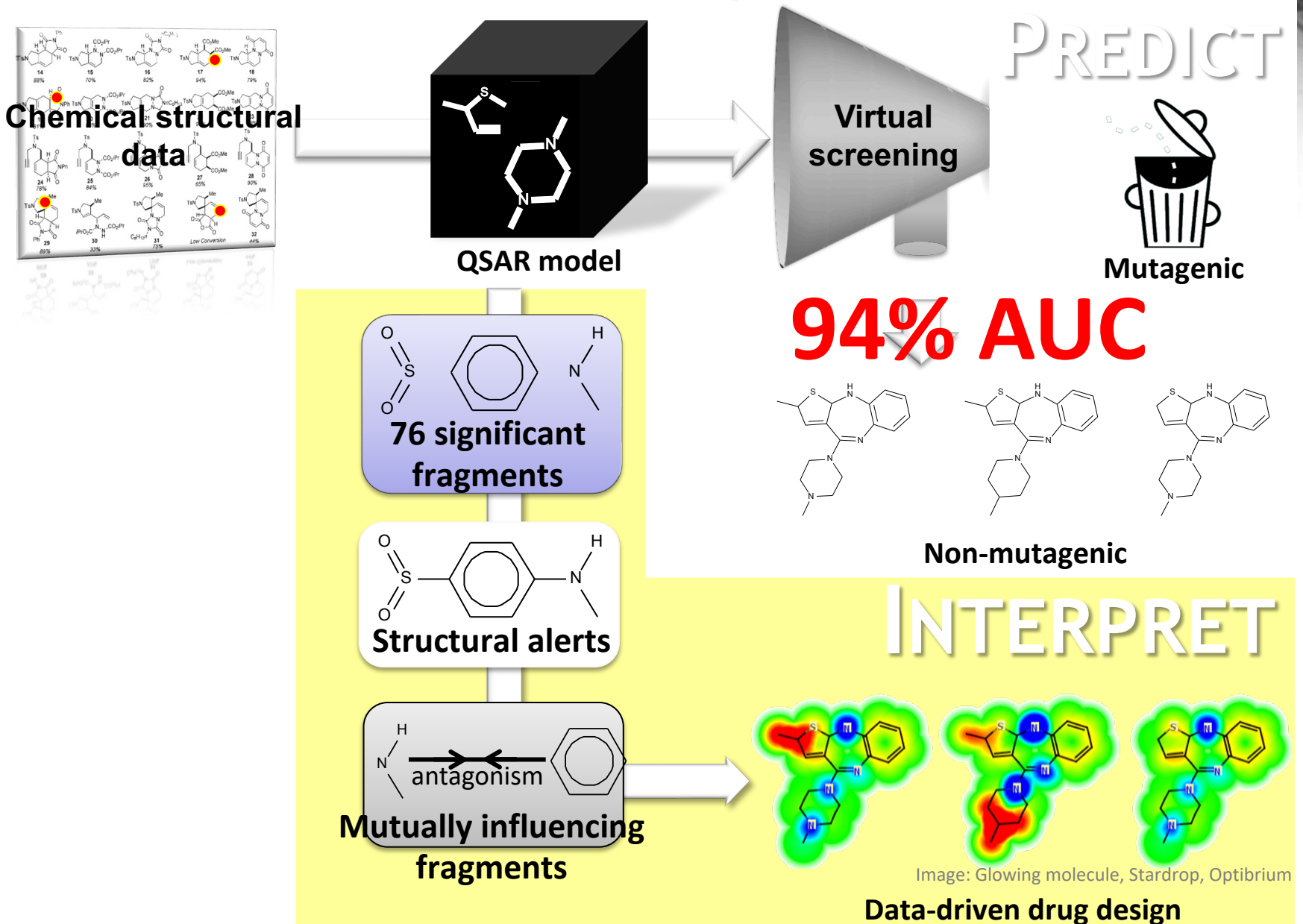
nitroalkanes (primary)
Nitro(prop – hex)ane
Non-mutagenic

Mechanism



Benigni 2011 *Chem Rev*
Helguera 2006 *Toxicol*
McCalla 1983 *Env Mutagen*

Marrying interpretability and statistical prediction accuracy: use QSAR models to validate descriptor-based assertions



Emerging applications of AI to chemical design and synthesis

THE ROYAL SOCIETY OF CHEMISTRY [GB] | https://www.chemistryworld.com/news/wanted-synthetic-chemists-humans-need-not-apply/3008401.article

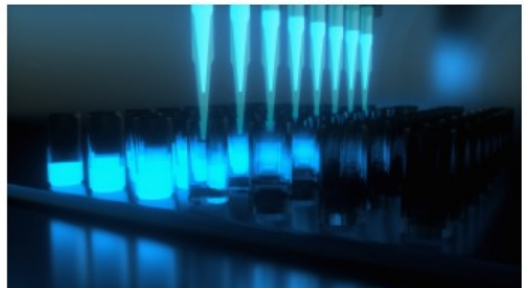
ToxPiWizard | ToxCast | file:///C:/Users/atrop | Google | New Tab | Save to Mendeley | Savings & Investment | wtharvey.com/m8n2 | Сериал Лист ожидан | GS_5

NEWS OPINION MATTER ENERGY EARTH LIFE CULTURE CAREERS PODCASTS WEBINARS LONG READS

NEWS

A brave new world of robot chemists and 'synthesiser farms' awaits

BY KATRINA KRÄMER | 24 JANUARY 2018



NEWS

Wanted: synthetic chemists (humans need not apply)

24 JANUARY 2018

Automation could free chemists from tedious lab work – if

Algorithm decides on chemical compromises when optimising self-driving experiments

BY HANNAH KERR | 3 SEPTEMBER 2018

F Häse, L Roch, and A Aspuru-Guzik, *Chimera: enabling hierarchy based multi-objective optimization for self-driving laboratories. Chem. Sci.*, 2018, DOI: 10.1039/c8sc02239a

QSAR Modeling: Going Deep

Deep Learning has (re)emerged as powerful ML algorithm.

- Higher predictivity than other algorithms such as RF and SVM.
  - "We found (1) that deep learning methods significantly outperform all competing methods" – Hochreiter group on ChEMBL data[1]
  - "Our results also show that models built with Deep Neural Networks had higher accuracy than those developed with simple machine learning algorithms" – **Tropsha group**, Tox21 Challenge[2]

Deep Learning does <u>not</u> always provide "deep" improvement

- Acute Toxicity: "Overall performance of DNN models on datasets of up to 30K compounds was similar to that of random forest (RF) models"[3]
- Bioactivity: "DNN achieved on average MCC units of 0.009 higher than SVM"[4]

Thinking Deep

"Although the performance of DNNs is generally better than RF using the standard DNN parameter settings, their predictive capability is variable under different parameter settings"[5]

1) DOI: 10.1039/C8SC00148K 2) DOI: 10.3389/fenvs.2016.00003
3) DOI: 10.1093/toxsci/kfy111 4) DOI: 10.1186/s13321-017-0226-y
5) DOI: 10.1021/ci500747n

Do newer methods such as Deep Learning truly always outperform other ML approaches?

Chemical Science

Large-scale comparison of machine learning methods for drug target prediction on ChEMBL†

Andreas Mayr, ‡[a] Günter Klambauer, ‡[a] Thomas Unterthiner, ‡[a] Marvin Steijaert,[b] Jörg K. Wegner,[c] Hugo Ceulemans,[c] Djork-Arné Clevert[d] and Sepp Hochreiter[a]

Cite this: *Chem. Sci.*, 2018, **9**, 5441

Authors' statement: "We found that deep learning methods significantly outperform all competing methods."

Table 1 Performance comparison of target prediction methods. The table input types. Overall, FNNs (second column) performed best. They signific representations of compounds and SmilesLSTM uses the SMILES represer

	FNN	SVM	RF
StaticF	0.687 ± 0.131	0.668 ± 0.128	0.665 ± 0.125
SemiF	0.743 ± 0.124	0.704 ± 0.128	0.701 ± 0.119
ECFP6	0.724 ± 0.125	0.715 ± 0.127	0.679 ± 0.128
DFS8	0.707 ± 0.129	0.693 ± 0.128	0.689 ± 0.120
ECFP6 + ToxF	0.731 ± 0.126	0.722 ± 0.126	0.711 ± 0.131
Graph			
SMILES			

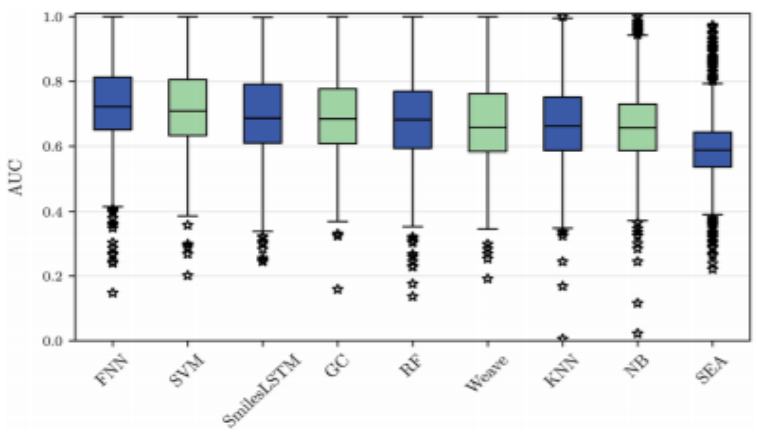


Fig. 2 Performance comparison of drug target prediction methods. The assay-AUC values for various target prediction algorithms based on ECFP6 features, graphs and sequences are displayed as boxplot.

Observation: the largest performance difference (AUC) between DNN and SVM or RF using the same descriptors is 0.04 (mind that SE is an order of magnitude larger, 0.12)!

Recent hype about chemical toxicity prediction

SCIENTIFIC AMERICAN

English

REACHAcross™

Home / How it Works

Alternative Methods, which shall identify such shortcomings. The methods are made practically available in a collaboration with Underwriters Laboratories (UL) (as REACHacross https://www.ulreachacross.com; last accessed June 30, 2018; and the UL Cheminformatics Suite, respectively).

Pricing

The cost is $295 per end-point, per substance.

With the REACHAcross™ database constantly evolving with the addition of new data sources, you have the ability for the $295 purchase price to re-generate your report for one year from the purchase date.

For quantity pricing, please call 518-640-9283

Print

Featured | Last comments | Popular

New evidence supports the hypothesis that

OXFORD UNIVERSITY PRESS | Toxicological Sciences

Oy Vey! A Comment on "Machine Learning of Toxicological Big Data Enables Read-Across Structure Activity Relationships Outperforming Animal Test Reproducibility".

Alves et al, Toxicol Sci. 2019 Jan 1;167(1):3-4

- Failure to take account of data heterogeneity
  - **Use predicted data to build the models**

These calls are made on the basis of OECD guideline studies, read across studies, QSAR studies and other information available in chemical dossiers submitted in service of REACH legisla-

- Use of inadequate data / Replication of compounds in a dataset
  - **No curation reported**
  - **"Not reliable" data present on ECHA database (major source of data)**

Future optimizations of the approach beside the expansion and curation of the database should address the similarity metrics employed (Luechtefeld and Hartung, 2017) and validate predic-

- Misuse/misinterpretation of statistics / Over-fitting of data / Failure to validate a QSPR correctly
  - **Use of compounds with conflicted annotation**
  - **Poor comparison of models with experimental assays**

For the 6 tests often referred to as "toxicological 6-pack" a reproducibility sensitivity of on average 70% was found (Table 2); the Simple RASAR matched this with on average the same 70%; by data fusion, 89% average sensitivity was achieved

…and the response…

MML UNC.EDU

Missing the Difference Between Big Data and Artificial Intelligence in RASAR Versus Traditional QSAR FREE

Thomas Luechtefeld, Dan Marsh, Thomas Hartung

Toxicological Sciences, Volume 167, Issue 1, January 2019, Pages 4–5,
https://doi.org/10.1093/toxsci/kfy287
Published: 30 November 2018

The letter challenges the approach as one would challenge a traditional QSAR, by which it ignores many attributes and consequences of the RASARs construction and performance as an implementation of big data and artificial intelligence (machine learning) (Hartung, 2016; Luechtefeld and Hartung, 2017).

To state it simply: the RASAR models are not traditional QSARs, wherein a highly curated, small training dataset is used to predict a single property based on chemical descriptors, ie, classifications per hazard. The published model uses data on 100 000+ chemical structures, calculates 5 billion+ similarities, and simultaneously makes 190 000 predictions for nine hazards of toxic properties of chemicals: 87% are correct, which should raise the question what we got right, not what we got wrong?

A brief history of “new” broad spectrum antibiotic discovery

Cell

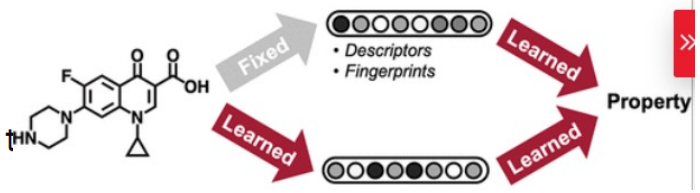
Volume 180, Issue 4, 20 February 2020, Pages 688-702.e13

zdnet.com/article/mits-deep-learning-found-an-antibiotic-for-a-germ-nothing-else-could-kill/

Analyzing Learned Molecular Representations for Property Prediction

Kevin Yang*, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay

Cite this: *J. Chem. Inf. Model.* 2019, 59, 8, 3370–3388
Publication Date: July 30, 2019



destroy a pathogen for which no cure has existed, and it could even help in t

well as previous graph neural architectures on both public and proprietary data sets. Our empirical findings indicate that while approaches based on these representations have yet to reach the level of experimental reproducibility, our proposed model nevertheless offers significant improvements over models currently used in industrial workflows.

A brief history of “new” broad spectrum antibiotic discovery

MML UNC.EDU

Correction to Analyzing Learned Molecular Representations for Property Prediction

Kevin Yang*, Kyle Swanson*, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay

Cite this: *J. Chem. Inf. Model.* 2019, 59, 12, 5304–5305
Publication Date: December 9, 2019

Due to an error in the processing of the random forest model’s predictions on classification data sets, our original random forest AUC numbers were incorrect on six public classification data sets—HIV, BACE, BBBP, Tox21, SIDER, and ClinTox—and on one proprietary classification data set—hPXR (class). We fixed the error

Table 1. (Random Split, Higher = Better) Comparison to Baselines on Public Datasets with Original and Fixed Random Forest Numbers Using a Random Split

data set	metric	D-MPNN	D-MPNN ensemble	RF on Morgan (original)	RF on Morgan (fixed)
HIV	ROC-AUC	0.816 ± 0.023	0.836 ± 0.020 (+2.40% $p = 0.01$)	0.641 ± 0.022 (−21.45%	0.819 ± 0.025 (+0.31% $p = 0.97$)
BACE	ROC-AUC	0.878 ± 0.032	0.898 ± 0.034 (+2.31% $p = 0.00$)	0.825 ± 0.039 (−6.08% p	0.898 ± 0.031 (+2.26% $p = 1.00$)
BBBP	ROC-AUC	0.913 ± 0.026	0.925 ± 0.036 (+1.23% $p = 0.01$)	0.788 ± 0.038 (−13.77%	0.909 ± 0.028 (−0.42% $p = 0.19$)
Tox21	ROC-AUC	0.845 ± 0.015	0.861 ± 0.012 (+1.95% $p = 0.00$)	0.619 ± 0.015 (−26.75%	0.819 ± 0.017 (−3.06% $p = 0.00$)
SIDER	ROC-AUC	0.646 ± 0.016	0.664 ± 0.021 (+2.79% $p = 0.01$)	0.572 ± 0.007 (−11.38%	0.687 ± 0.014 (+6.35% $p = 1.00$)
ClinTox	ROC-AUC	0.894 ± 0.027	0.906 ± 0.043 (+1.33% $p = 0.05$)	0.544 ± 0.031 (−39.13%	0.759 ± 0.060 (−15.12% $p = 0.00$)

A brief history of “new” broad spectrum antibiotic discovery

MML UNC.EDU

c&en CHEMICAL & ENGINEERING NEWS

TOPICS MAGAZINE COLLECTIONS VIDEOS JOBS

COMPUTATIONAL CHEMISTRY

AI finds molecules that kill bacteria, but would they make good antibiotics?

Experts praise the approach while remaining skeptical that the highlighted molecules could reach the clinic

CORRECTION

This story was updated on March 5, 2020, to include information about a previous study that identified antibiotic activity for halicin.

https://cen.acs.org/physical-chemistry/computational-chemistry/AI-finds-molecules-kill-bacteria/98/web/2020/02?utm_source=Twitter&utm_medium=Social&utm_campaign=CEN

AI: words of warning

THE CONVERSATION

How big da a big crisis in science

December 13, 2018 6.44am EST

BBC Weather Shop Reel

NEWS

AAAS ence

crisis'

urban DICTIONARY Browse Vote Store

TOP DEFINITION

Ai Ai Ai

Ai Ai Ai is a **phrase** used by most **evil** people with **bad intentions**

- Machi ly to find interes gh huge amour
- Machi to analyse data a pletely wrong
- There is general recognition of a reproducibility crisis in science …

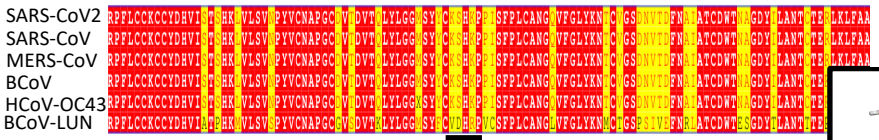
Dr Genevera Allen, Rice University, AAAS meeting, feb 2019

Tight Integration of Computational tools and experiment

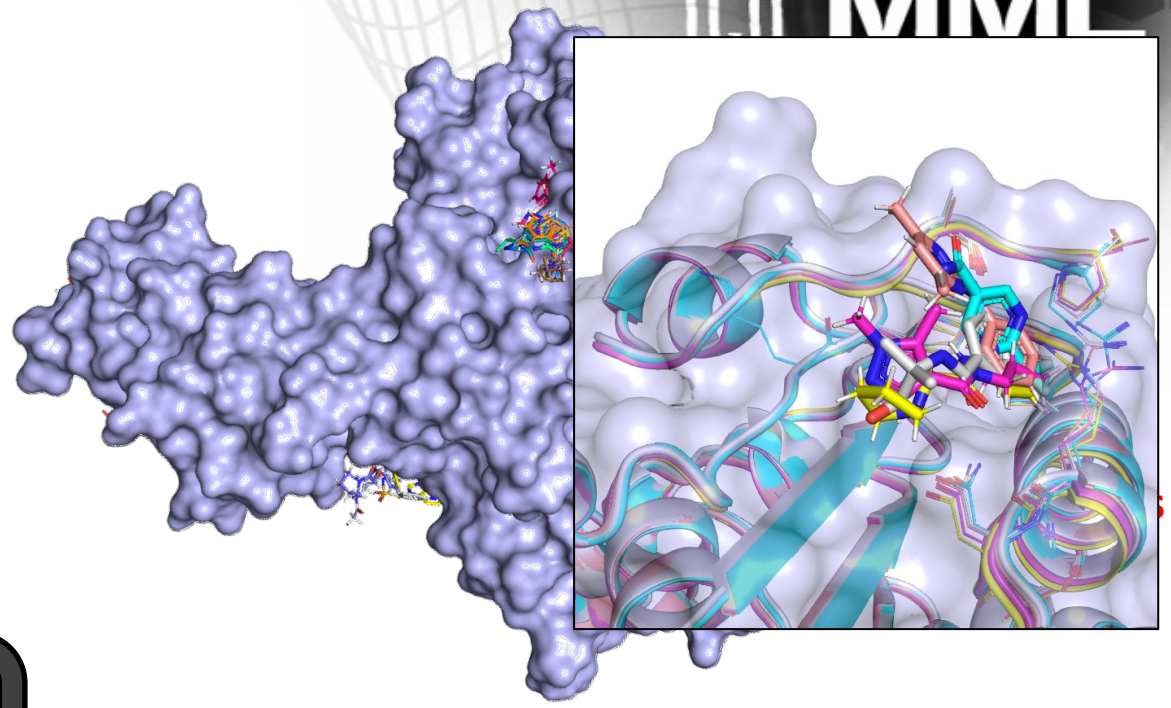
Target: NSP13 (project 2)

VGACVLCNSQTSLRCGACIRRPFLCCKCCYDHVISTSHKLVLSVNPYVCNAPGCDVTDVTQLYLG
GMSYYCKSHKPPISFPLCANGQVFGLYKNTCVGSDNVTDFNAIATCDWTNAGDYILANTCTER
LKLFAAETLKATEETFKLSYGIATVREVLSDRELHLSWEVGKPRPPLNRNYVFTGYRVTKNSKVQI
GEYTFEKGAVVYRGTTTYKLNVGDYFVLTSHTVMPLSAPTLVPQEHYVRITGLYPTLNISDEFSSN
VANY…

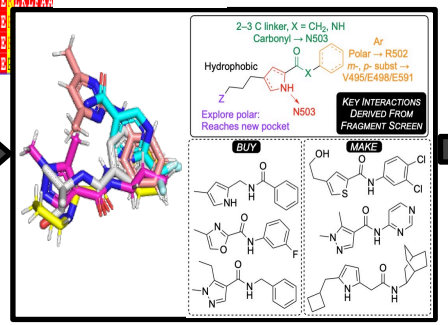
ENDscript2/Blast



Knowledge mining
(e.g., PDB, PubMed)
(AlphaFold2)

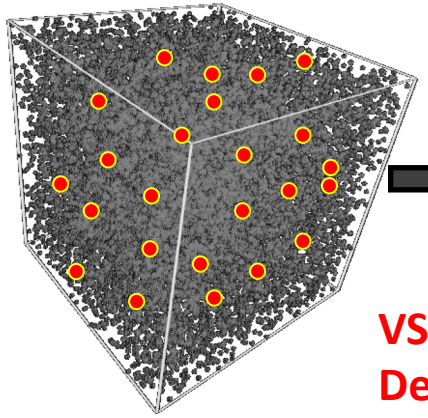


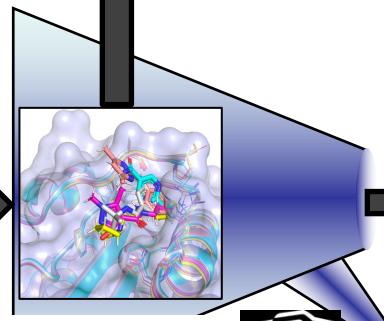
Design



Maestro/PyMol

Purchasable library





VS: Glide, DeepDocking SimSearch

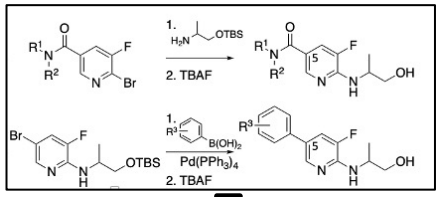
Predicted Actives

Predicted inactives

Screening/ Testing

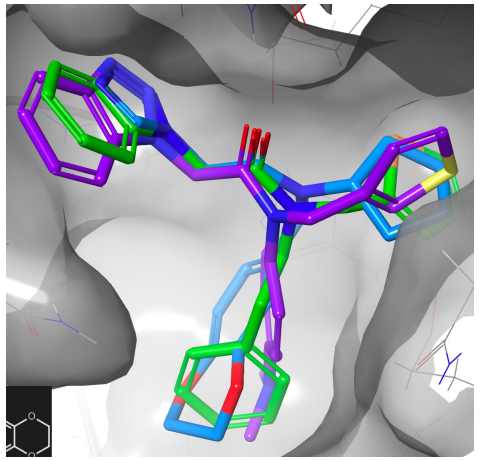
Hit-to-Lead

QSAR/MD-FES

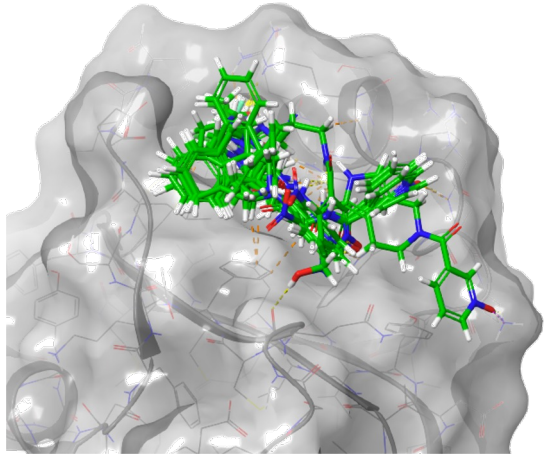


Enamine REAL Space (~38B) virtual screenings for AViDD targets

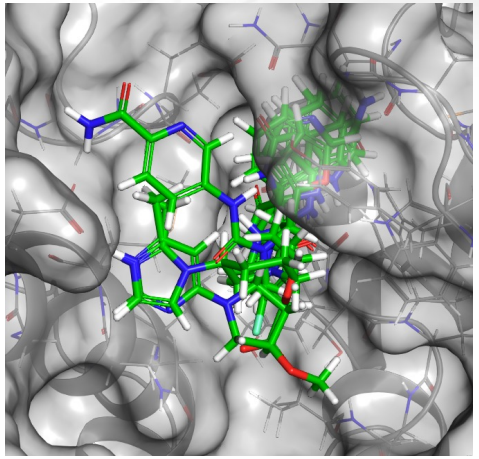
SARS-CoV2 Mpro



SARS-CoV2 Nsp13



CHIKV nsp2-protease



Nominations:

- 150 compounds have been purchased
- **1 compound showed high nM activity**
- **7 compounds are in ~10uM range**

Nominations:

- 50 compounds have been purchased
- 30 de novo generated compounds being synthesized
- **3 compounds showed < 10uM activity**

Nominations:

- 150 purchasable compounds

The HIDDEN GEM workflow is currently being executed for multiple viral targets

Societal issues: how to improve the quality of published data and models

- Develop clear guidance (raise acceptance bar) for both authors and reviewers
  - Minimal model acceptance criteria similar to JMC requiring data on compound composition and purity
  - Availability of both curated data and models similar to protein journals requiring deposit to PDB to accept a paper describing new protein structure
- Inform applied journals about our acceptance rules
- Work with data journals and database groups (e.g., ChEMBL, PubChem) on data quality standards
- Publish in high-profile journals

Guidelines and associated software tools for reporting, storing, and sharing detailed information considered to be important to include with published data sets on bioactive entities:

NATURE REVIEWS | DRUG DISCOVERY
VOLUME 10 | SEPTEMBER 2011 | 661

OPINION

PERSPECTIVES

Minimum information about a bioactive entity (MIABE)

Sandra Orchard, Bissan Al-Lazikani, Steve Bryant, Dominic Clark, Elizabeth Calder, Ian Dix, Ola Engkvist, Mark Forster, Anna Gaulton, Michael Gilson, Robert Glen, Martin Grigorov, Kim Hammond-Kosack, Lee Harland, Andrew Hopkins, Christopher Larminie, Nick Lynch, Romeena K. Mann, Peter Murray-Rust, Elena Lo Piparo, Christopher Southan, Christoph Steinbeck, David Wishart, Henning Hermjakob, John Overington and Janet Thornton

MML NC.EDU

- **A** Molecule properties (names, structure, InChi, salt, prodrug, …)
- **B** Molecule production (chemical synthesis, purity, characterization, …)
- **C** Physicochemical properties (molecular weight, water solubility, hydrophobicity, …)
- **D** In vitro cell-free assays (primary target, assay details and parameters, delivery systems, secondary gene targets, …)
- **E** Cellular assays (cell type, conditions, assay type, …)
- **F** Whole-organism studies (animal/plant studies, disease model, toxicology, DDI, …)
- **G** Pharmacokinetic studies *(absorption, dosing route, half-life, Vmax, metabolism, …)*

Conclusions and Outlook

- Rapid accumulation of large biomolecular datasets and VS libraries (especially, in public domain):
  - Strong need for both chemical and biological data curation
- Novel approaches towards Integration of inherent chemical properties with additional data streams
  - improve the outcome of structure – in vitro – in vivo extrapolation
- Interpretation of significant chemical and biological descriptors emerging from externally validated models
  - inform the selection or design of effective and safe chemicals
- Exciting developments at the interface between computational and organic chemistry
  - Critical shift from discovery in databases to design and AI-driven robotics (SDL!)
- Tool and data sharing
  - Pubic web portals (e.g., Chembench, OCHEM)