NC STATE UNIVERSITY

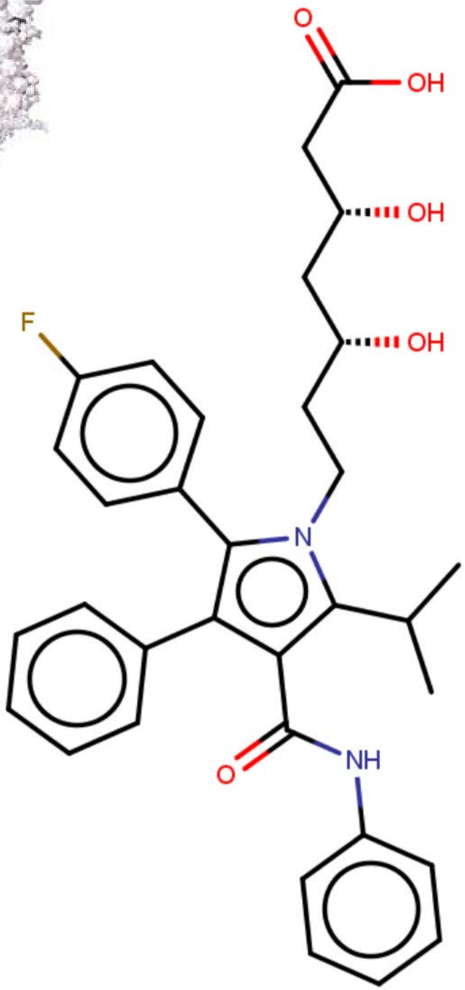
Characterizing, Navigating, and Modeling the Chemical Space Using Next-Generation Cheminformatics Methods

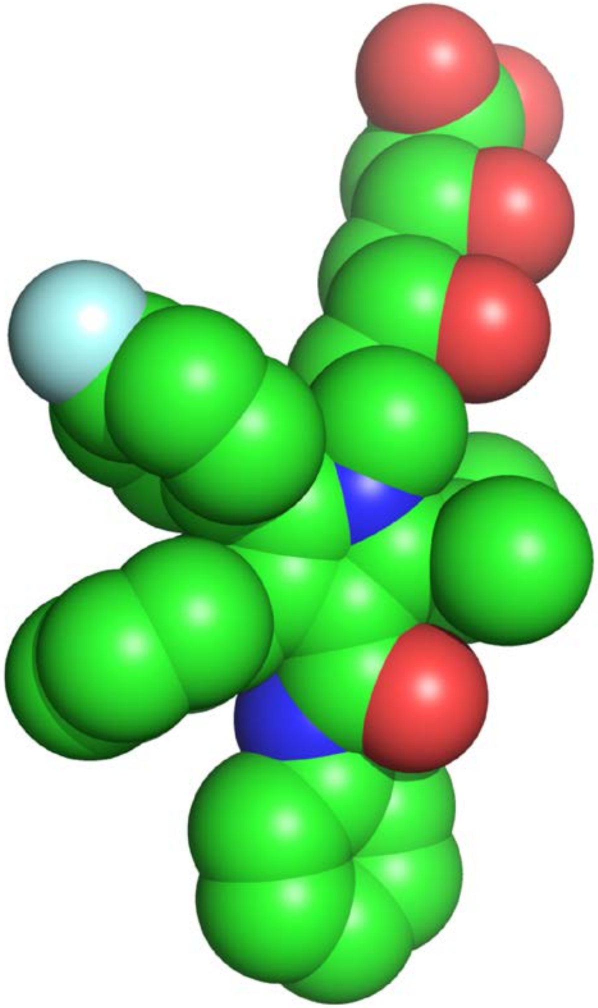
Denis Fourches, PhD

Department of Chemistry, Bioinformatics Research Center,
North Carolina State University, USA
www.fourches-laboratory.com

Follow on ResearchGate LinkedIn

$C_{33}H_{35}FN_2O_5$





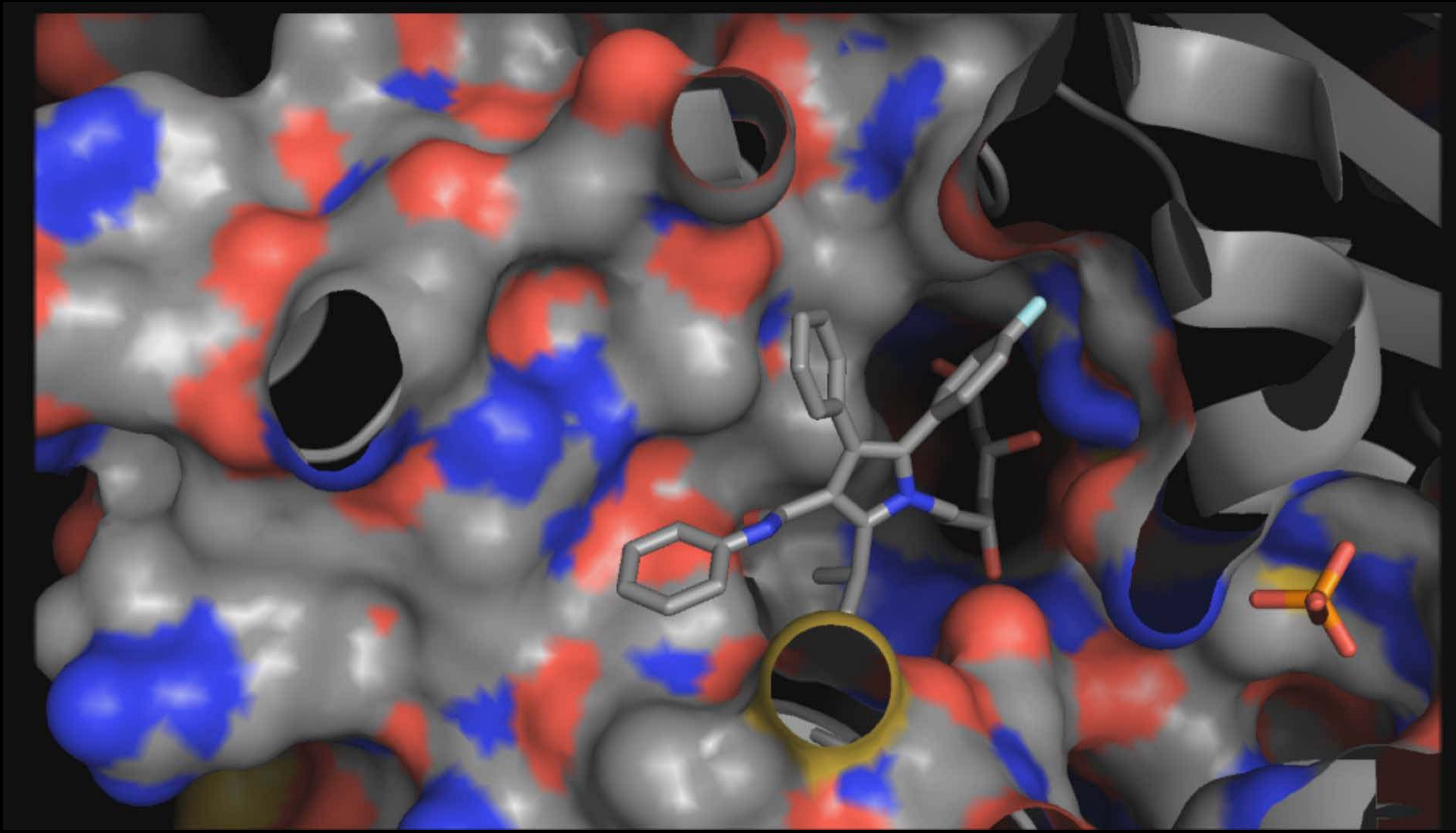
0D/1D

2D

3D

LIPITOR (atorvastatin)

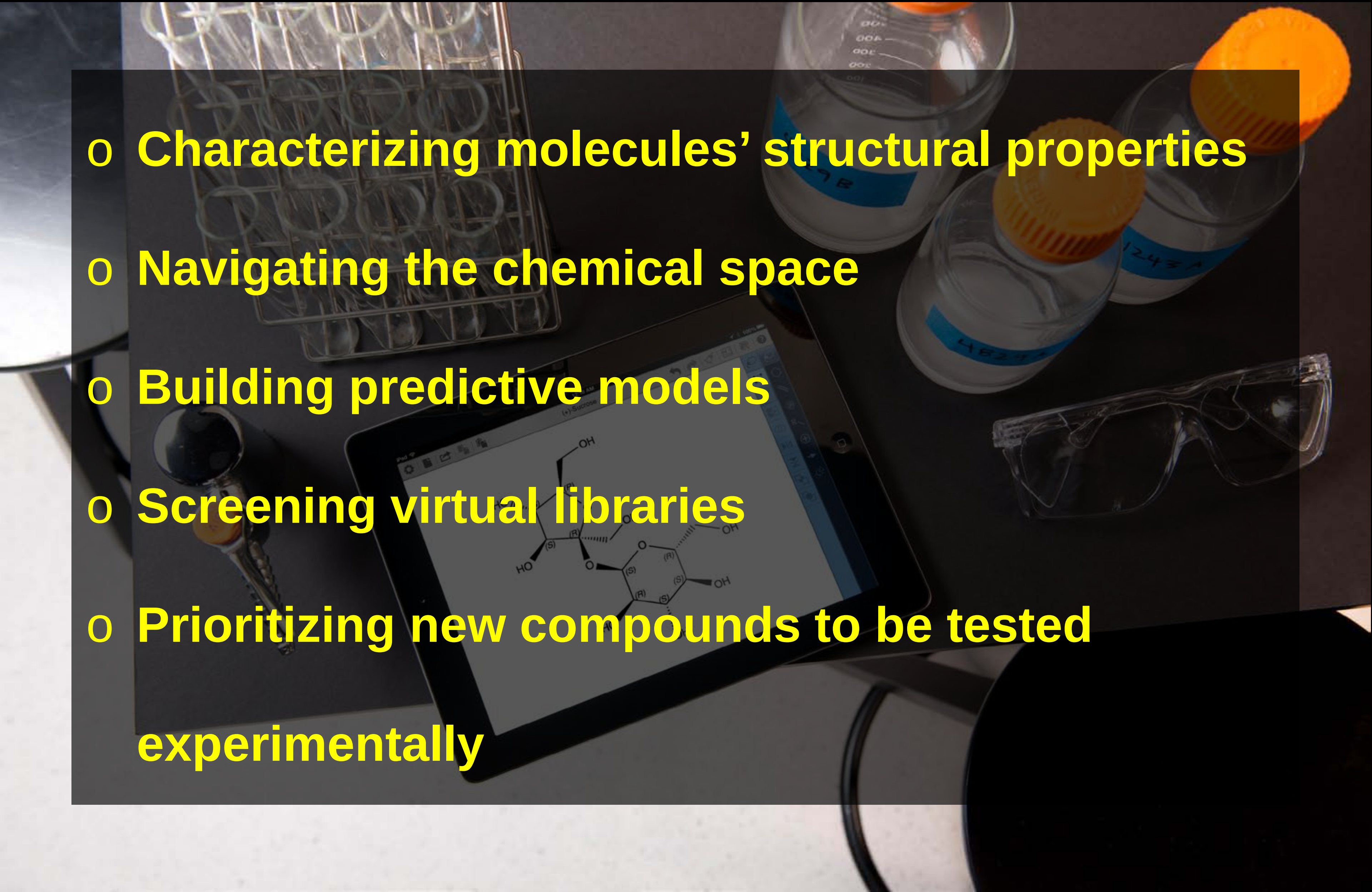
World's best-selling drug of all time ($125 billion over 15 years)



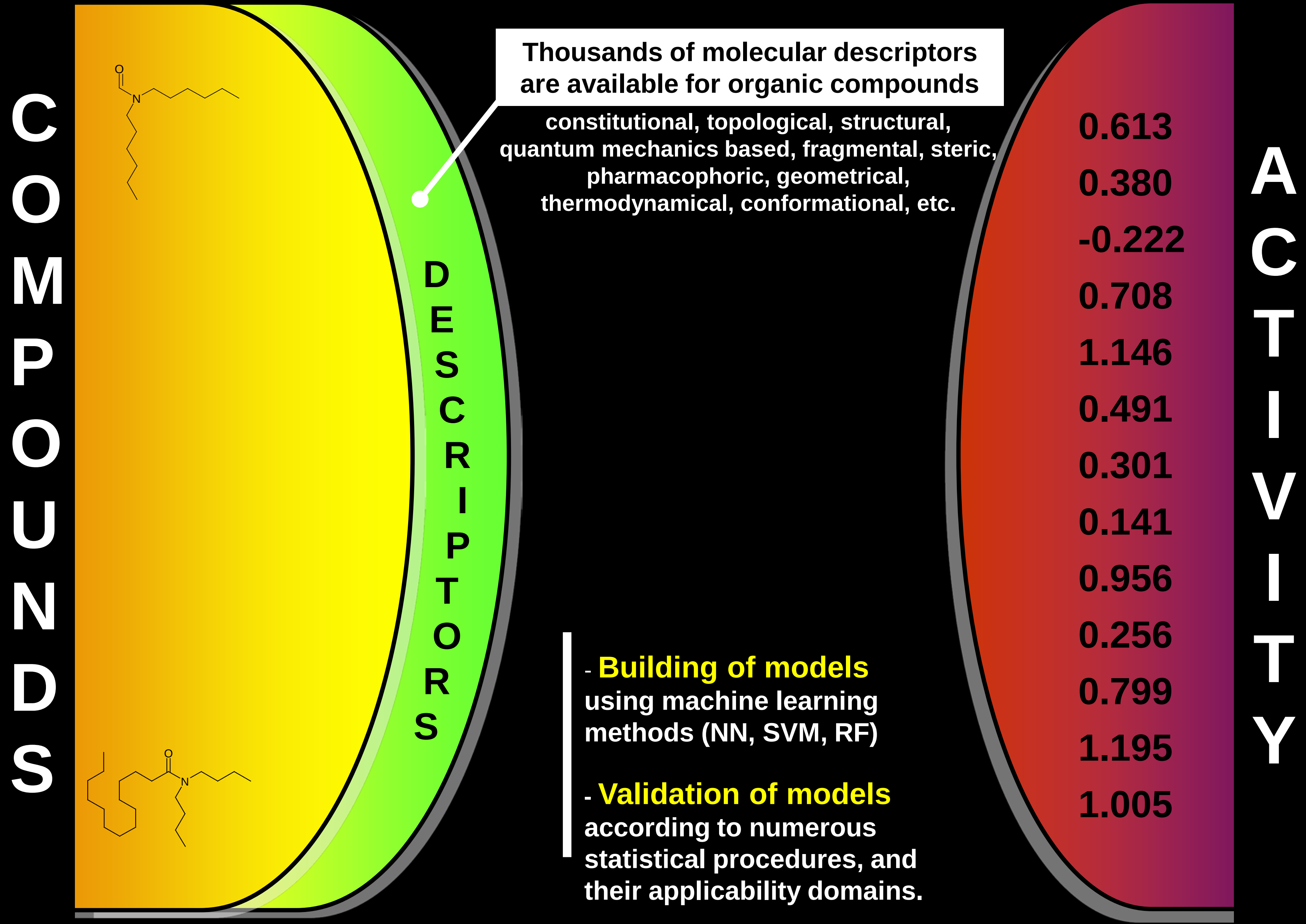
Competitive inhibitor of HMG-CoA reductase in liver → **Lower blood cholesterol**

PDB code = 1HWK

Cheminformatics is becoming an essential element in the chemist’s toolbox



- Characterizing molecules’ structural properties
- Navigating the chemical space
- Building predictive models
- Screening virtual libraries
- Prioritizing new compounds to be tested experimentally



Thousands of molecular descriptors are available for organic compounds

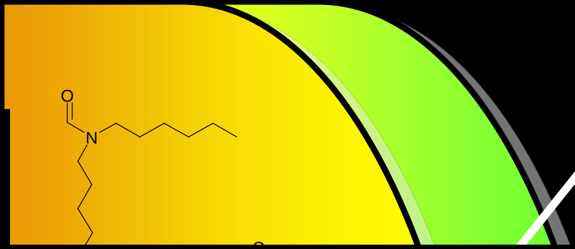
constitutional, topological, structural, quantum mechanics based, fragmental, steric, pharmacophoric, geometrical, thermodynamical, conformational, etc.

- **Building of models** using machine learning methods (NN, SVM, RF)

- **Validation of models** according to numerous statistical procedures, and their applicability domains.

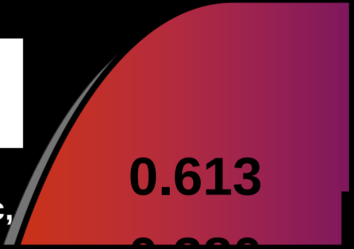
Cherkasov, Muratov, Fourches, et al. 2014, J Med Chem, 57(12), 4977-5010

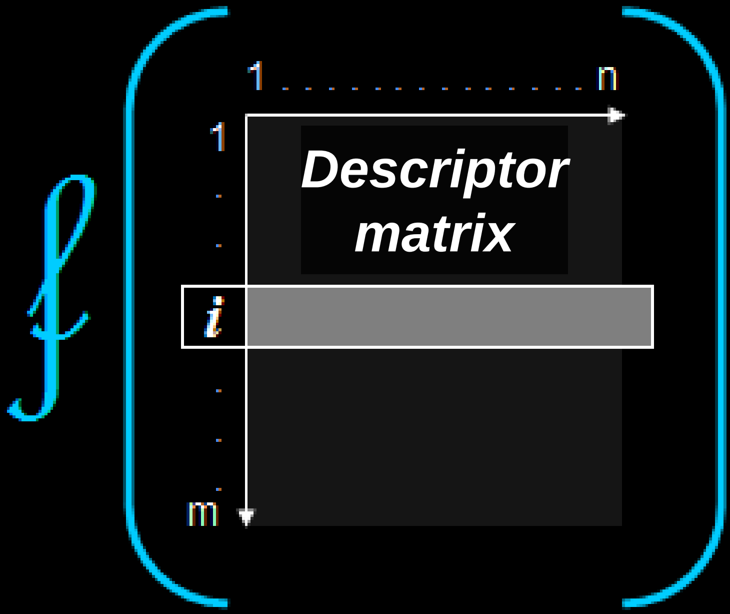
COMPOUNDS



Thousands of molecular descriptors are available for organic compounds

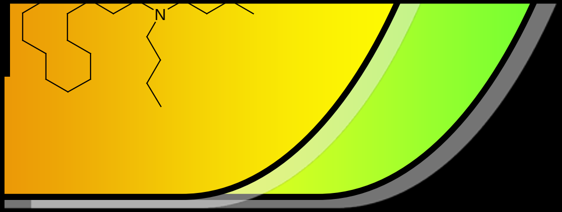
constitutional, topological, structural, quantum mechanics based, fragmental, steric, pharmacophoric, geometrical



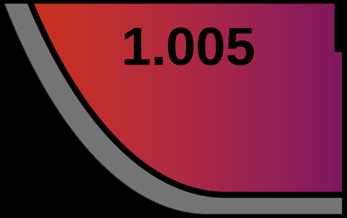


= ACTIVITY (*i*)

With m molecules and n descriptors

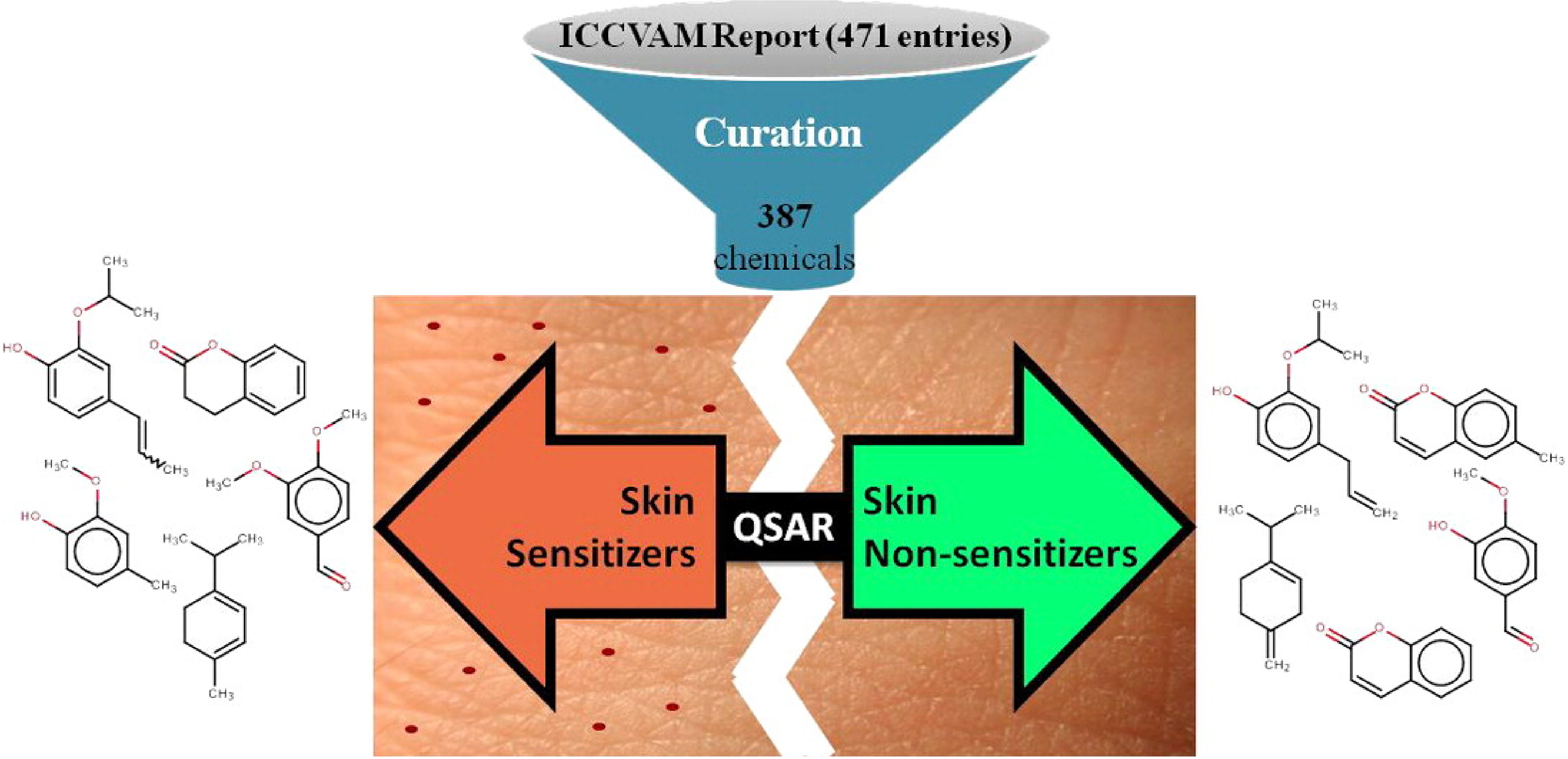


- **Validation of models** **according to numerous statistical procedures, and their applicability domains.**



ACTIVITY

Cherkasov, Muratov, Fourches, et al. 2014, J Med Chem, 57(12), 4977-5010



ICCVAM Report (471 entries)

Curation

387 chemicals

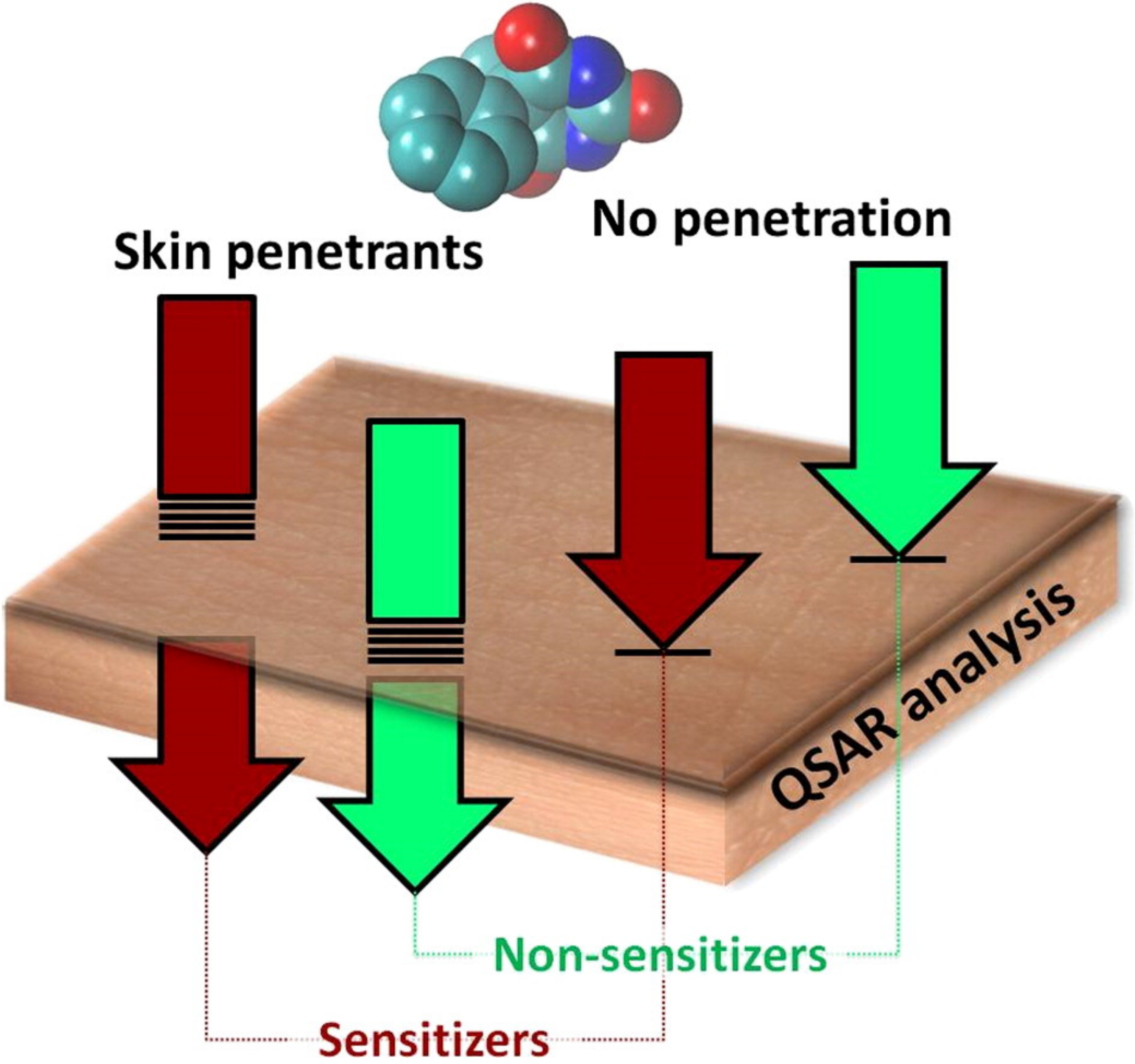
Skin Sensitizers

QSAR

Skin Non-sensitizers

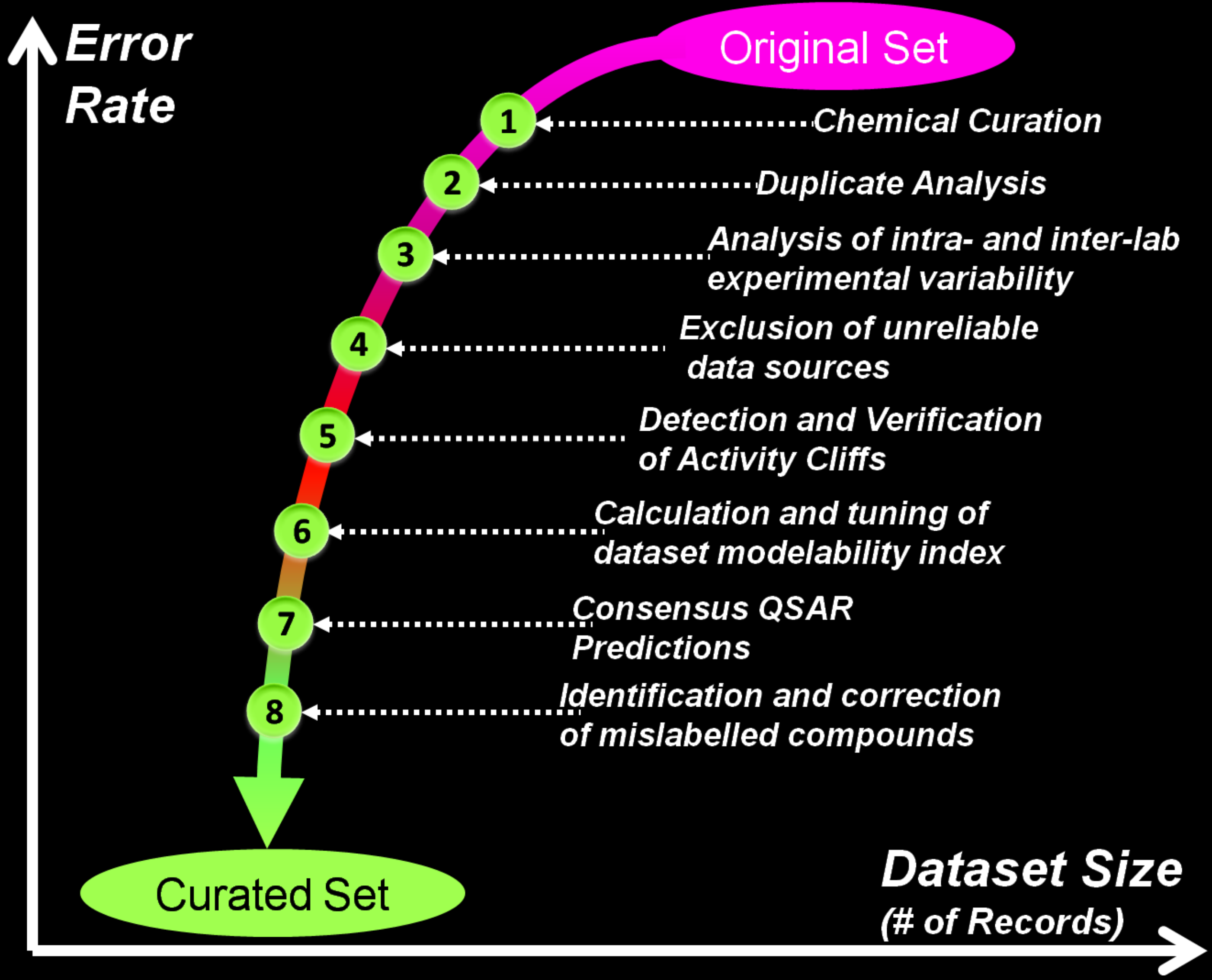
Correct Classification Rate (CCR) for QSAR models discriminating sensitizers from non-sensitizers was 71–88% when evaluated on several external validation sets, within a broad AD, with positive (for sensitizers) and negative (for non-sensitizers) predicted rates of 85% and 79% respectively.

Alves, Muratov, Fourches, Strickland, Kleinstreuer, Andrade, Tropsha. Toxicol Appl Pharmacol. 2015 Apr 15;284(2):262-72.



Alves, Muratov, Fourches, Strickland, Kleinstreuer, Andrade, Tropsha. Toxicol Appl Pharmacol. 2015 Apr 15;284(2):273-80.

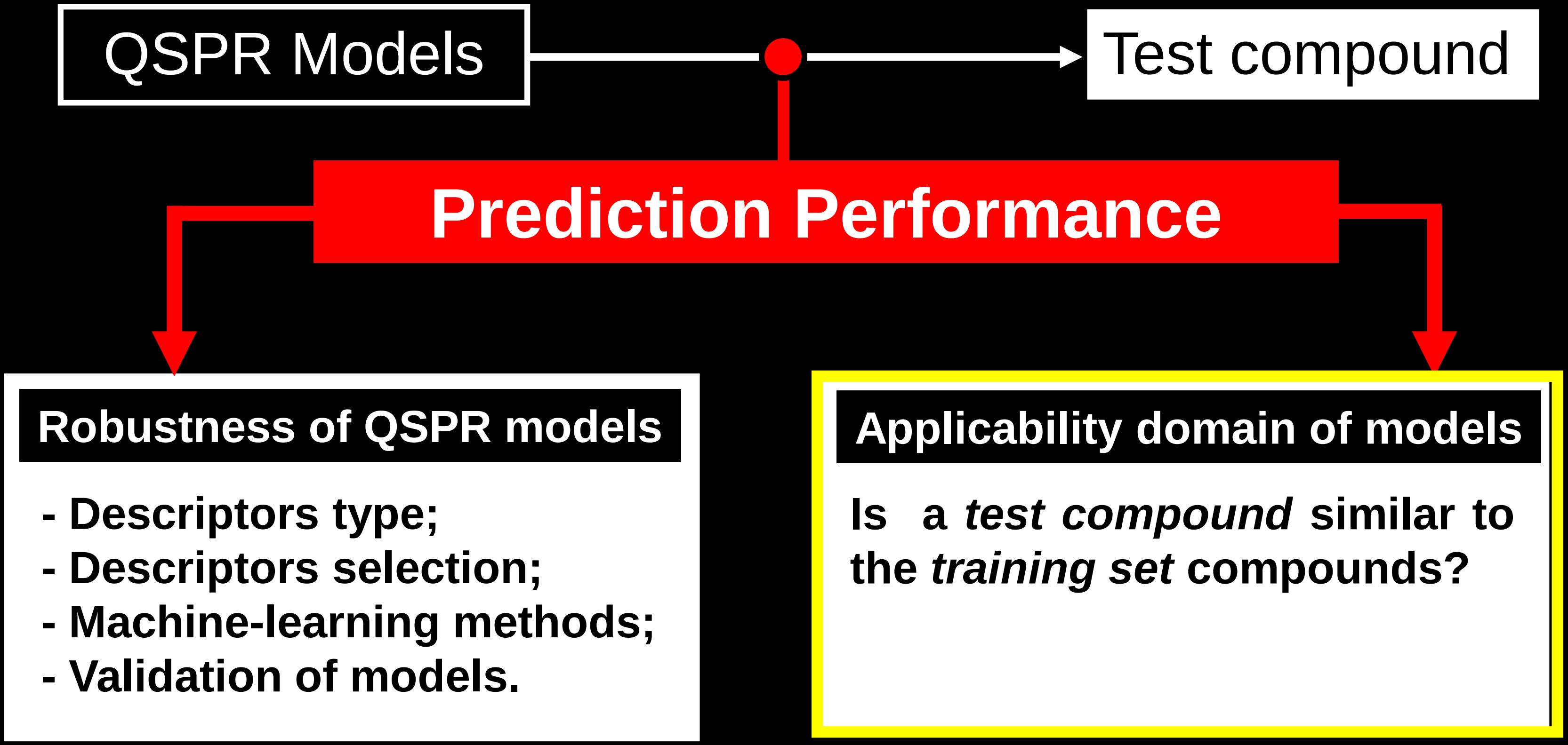
Curation of Chemogenomics Data


Error Rate
Original Set
1
Chemical Curation
2
Duplicate Analysis
3
Analysis of intra- and inter-lab experimental variability
4
Exclusion of unreliable data sources
5
Detection and Verification of Activity Cliffs
6
Calculation and tuning of dataset modelability index
7
Consensus QSAR Predictions
8
Identification and correction of mislabelled compounds
Curated Set
Dataset Size (# of Records)

Fourches, Muratov, Tropsha. Nature Chemical Biology. 2015, 11, 535.

Fourches, Muratov, Tropsha. JCIM. 2016, In Press.

Can any QSAR model, even if well validated, be applied to any molecule ?



QSPR Models

Test compound

Prediction Performance

Robustness of QSPR models

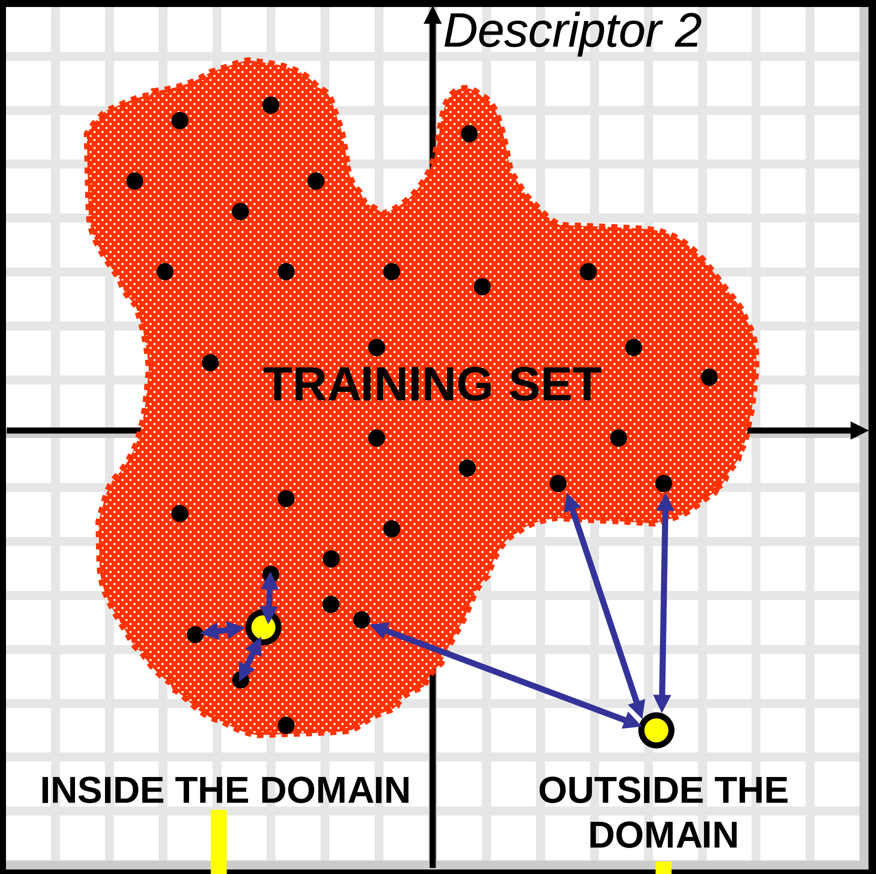
- **Descriptors type;**
- **Descriptors selection;**
- **Machine-learning methods;**
- **Validation of models.**

Applicability domain of models

Is a *test compound* similar to the *training set* compounds?

Applicability Domain of a given QSAR model

● = TEST COMPOUND



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

$$D_i \leq \langle D_k \rangle + Z \times s_k$$

with Z, an empirical parameter (0.5 by default)

Tropsha et al., J. Med. Chem, 2002, 45, p. 2811-2823

AD parameter of applicability domain

$$AD_i = D_i / (\langle D_k \rangle + Z \times s_k) \times \mathbf{100}$$

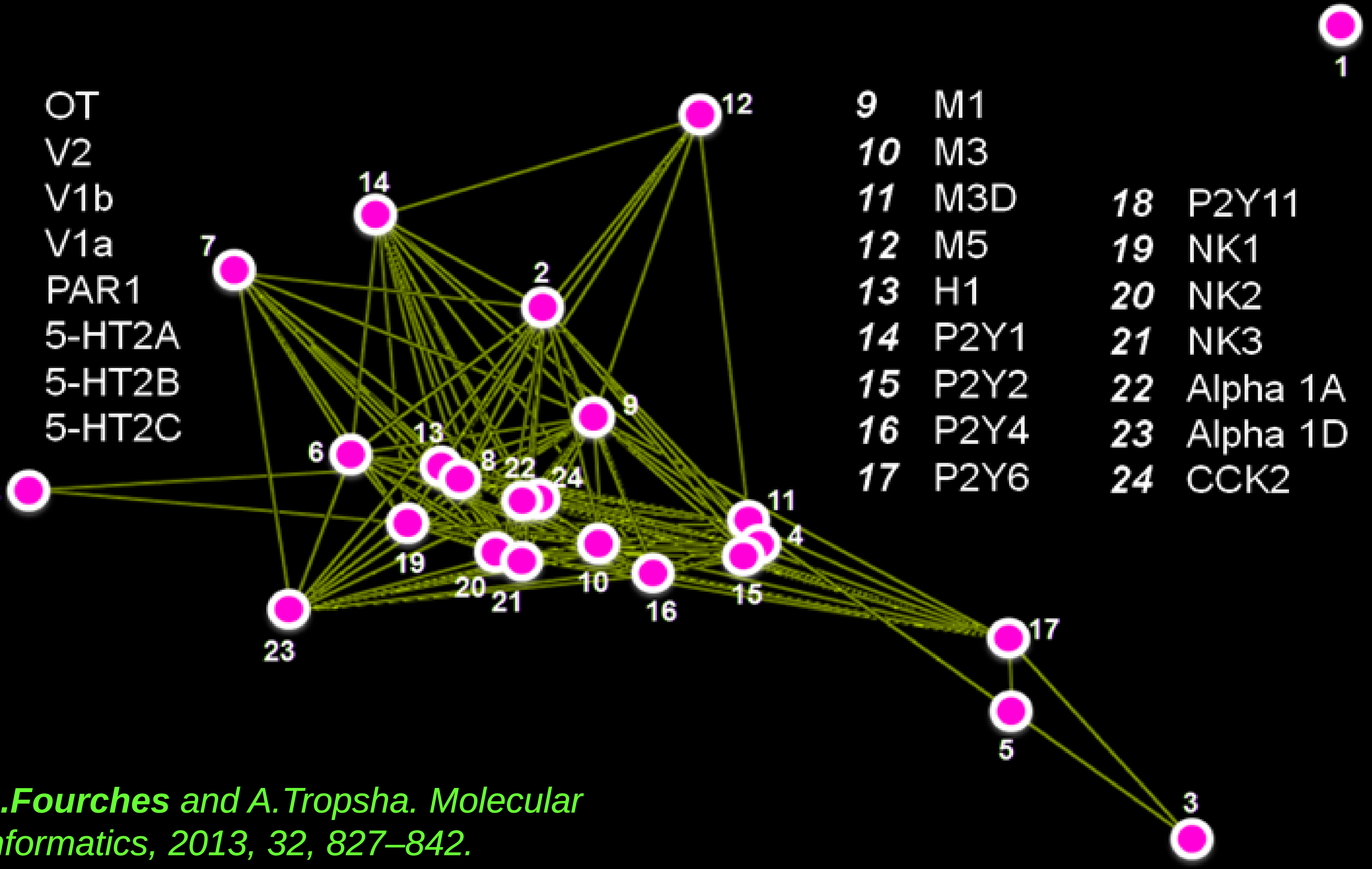
AD_i ≤ 100 %

Will be predicted by the model

AD_i > 100 %

Will not be predicted by the model

Development of New Computational Tools Adapted to Hyper-Dimensional HTS data to Analyze Drugs' Polypharmacology



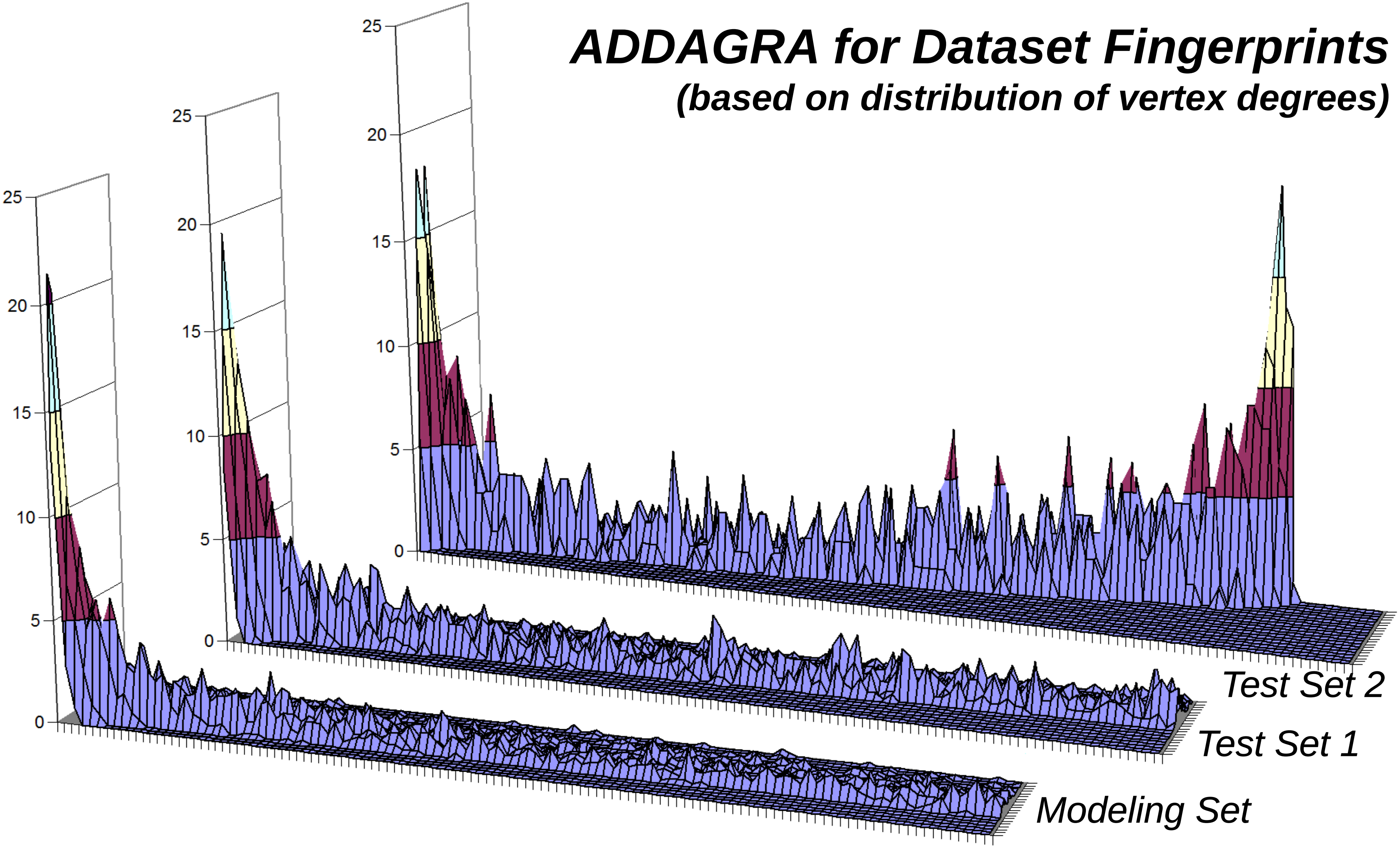
1	OT
2	V2
3	V1b
4	V1a
5	PAR1
6	5-HT2A
7	5-HT2B
8	5-HT2C
9	M1
10	M3
11	M3D
12	M5
13	H1
14	P2Y1
15	P2Y2
16	P2Y4
17	P2Y6
18	P2Y11
19	NK1
20	NK2
21	NK3
22	Alpha 1A
23	Alpha 1D
24	CCK2

***D.Fourches** and A.Tropsha. Molecular Informatics, 2013, 32, 827–842.*

Visualizing and comparing chemical datasets using the ADDAGRA approach

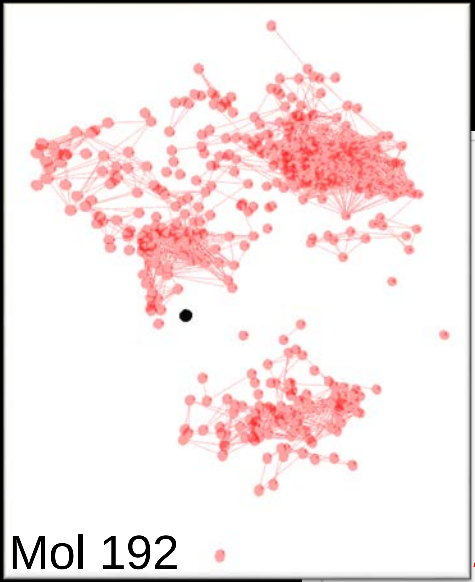
ADDAGRA for Dataset Fingerprints

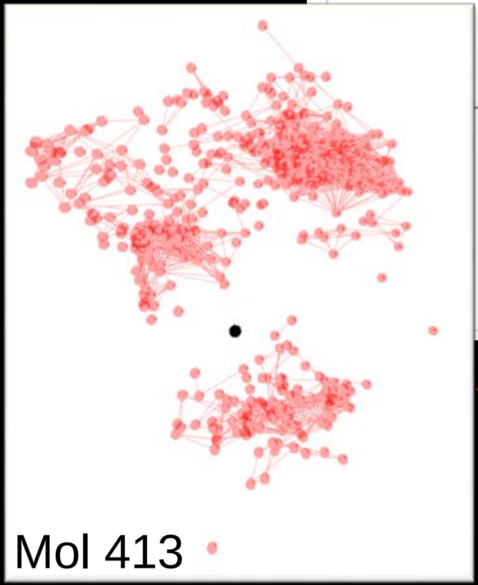
(based on distribution of vertex degrees)

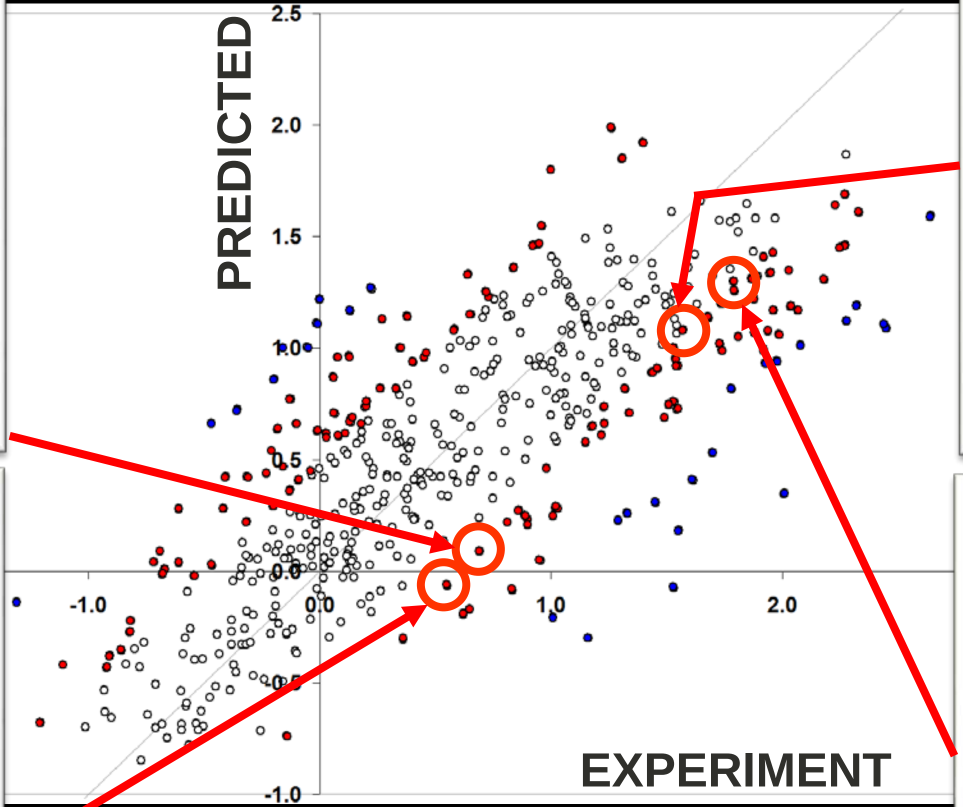


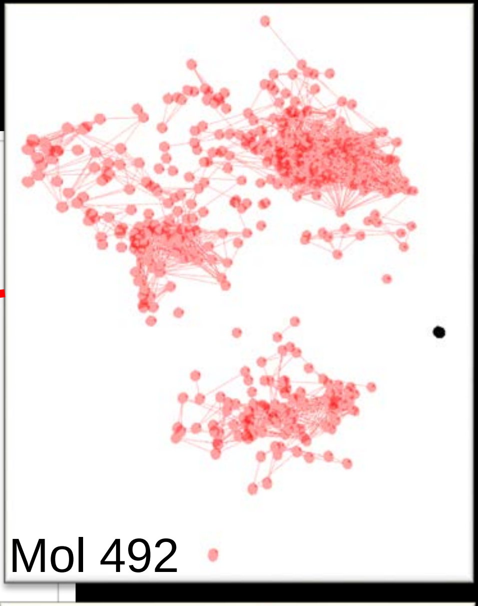
__D.Fourches__ and A.Tropsha. Using Graph Indices for the Analysis and Comparison of Chemical Datasets. Molecular Informatics, 2013, 32, 827–842.

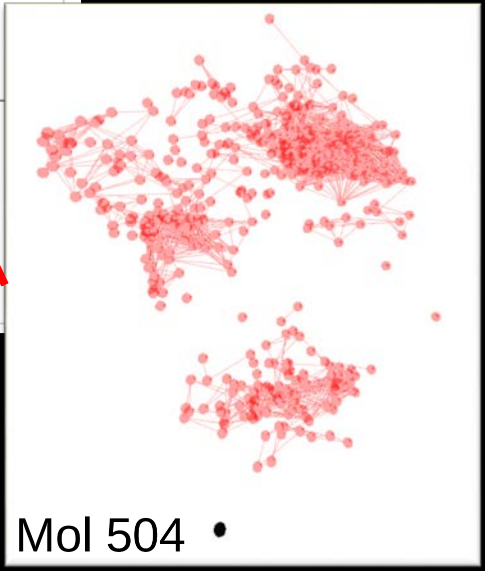
ADDAGRA for the analysis of prediction outliers







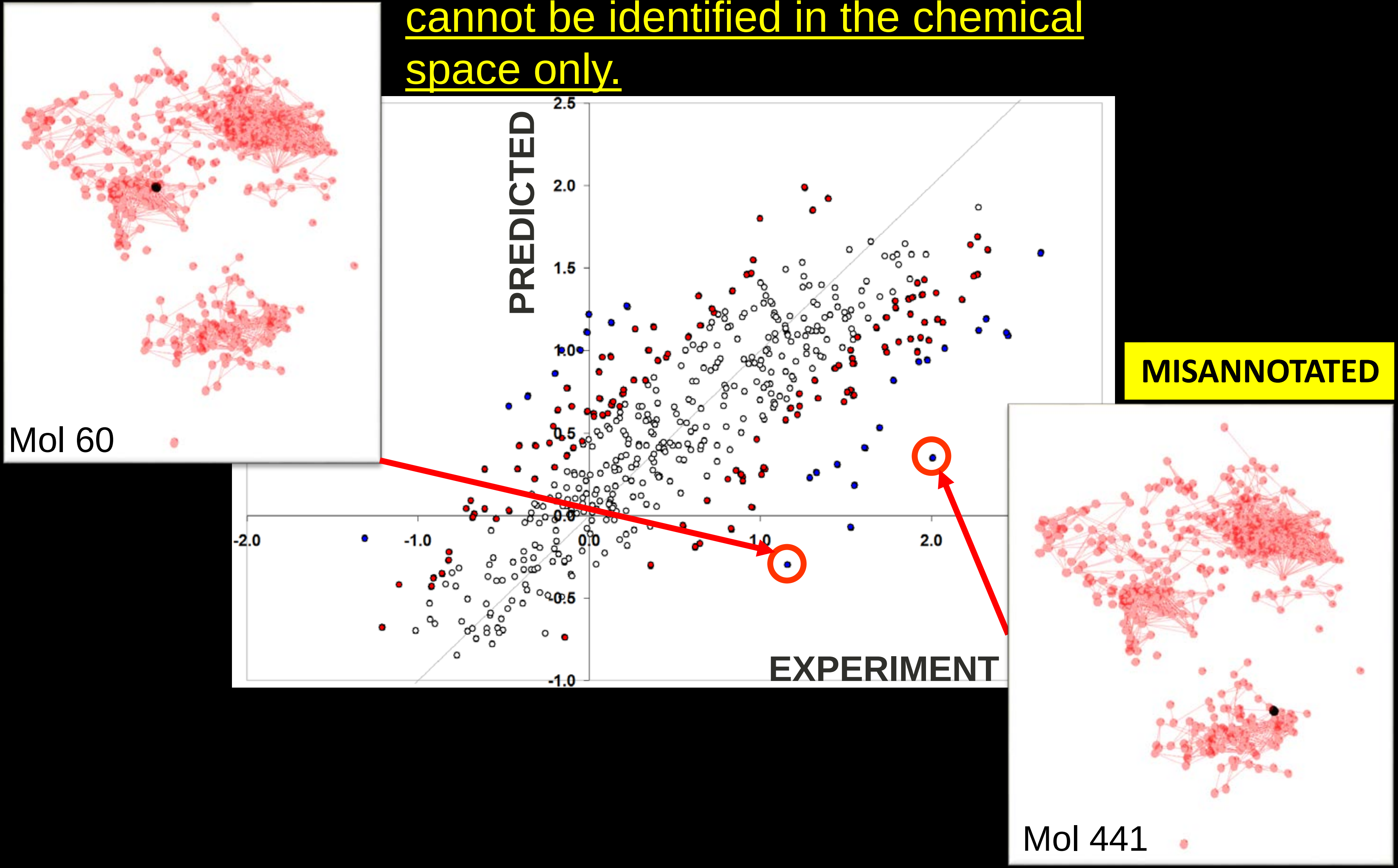




Many prediction outliers also correspond to outliers in the descriptor space.

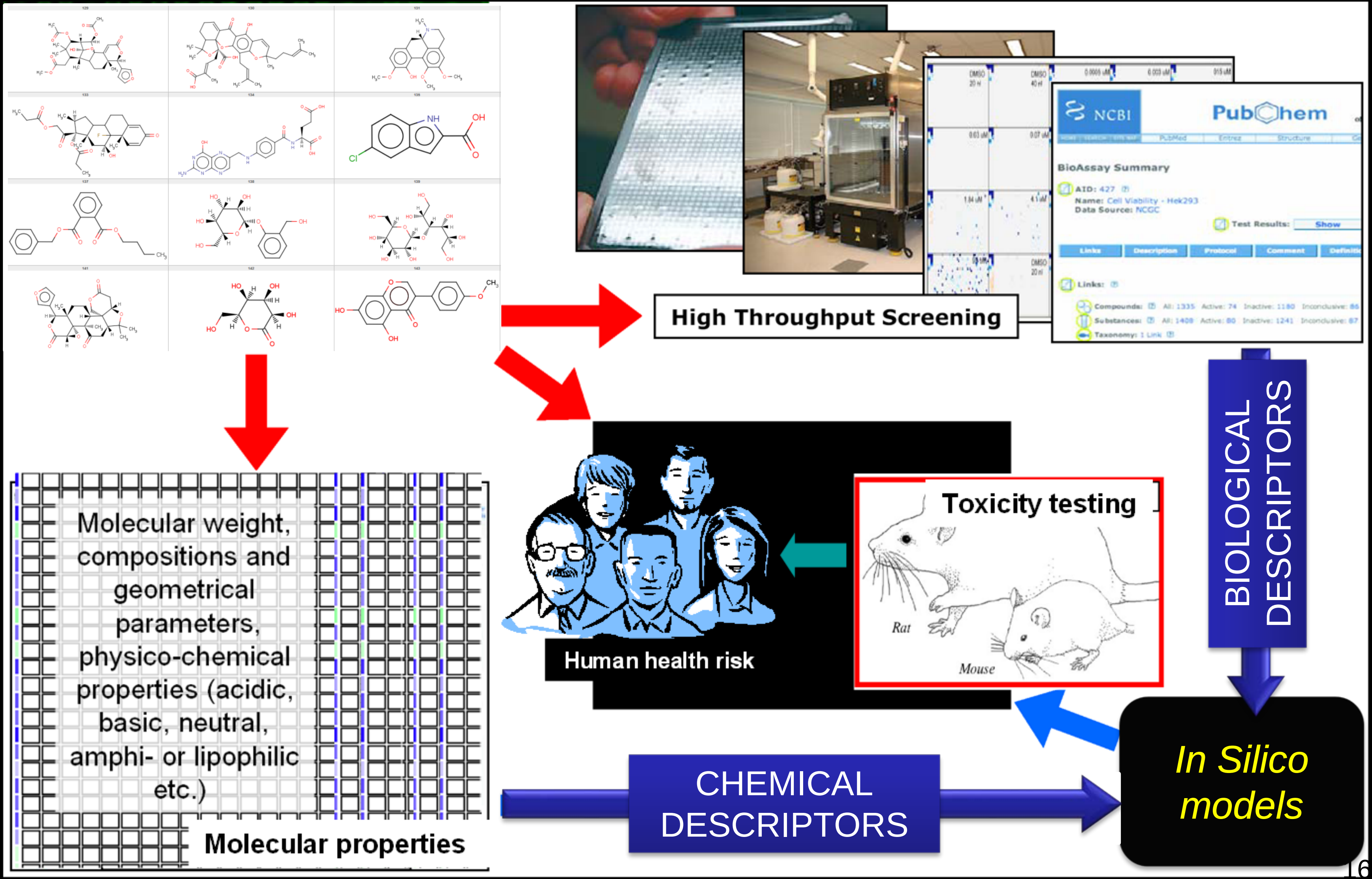
ADDAGRA for the analysis of prediction outliers

But still, some large outliers cannot be identified in the chemical space only.



Hybrid modeling using both chemical and biological descriptors

Environ Health Perspect. 2011,119(3):364-70
Chem Res Toxicol. 2011, 24(8):1251-62



High Throughput Screening

BIOLOGICAL DESCRIPTORS

Toxicity testing

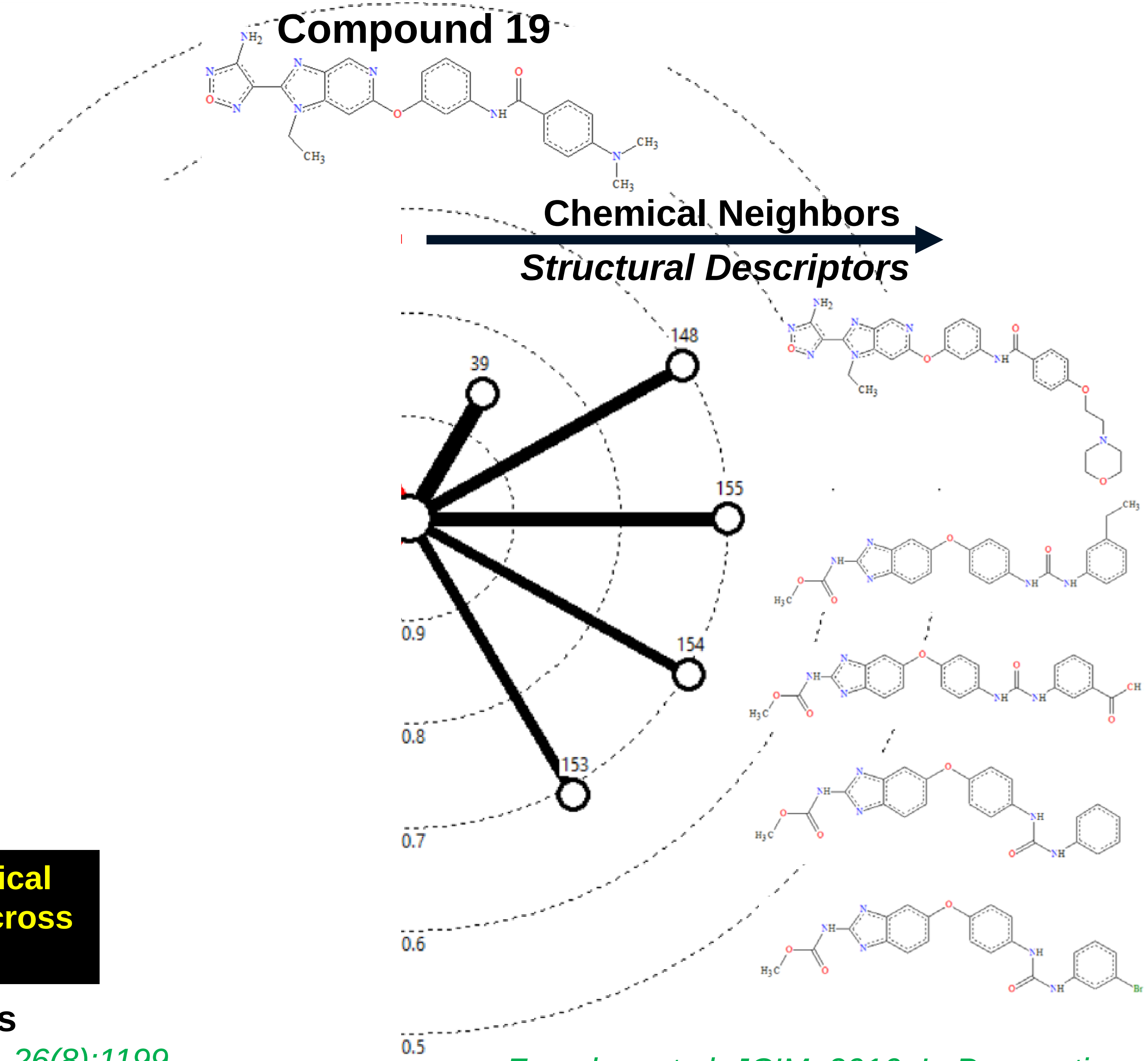
Human health risk

Molecular weight, compositions and geometrical parameters, physico-chemical properties (acidic, basic, neutral, amphi- or lipophilic etc.)

Molecular properties

CHEMICAL DESCRIPTORS

In Silico models



Compound 19

Chemical Neighbors

Structural Descriptors

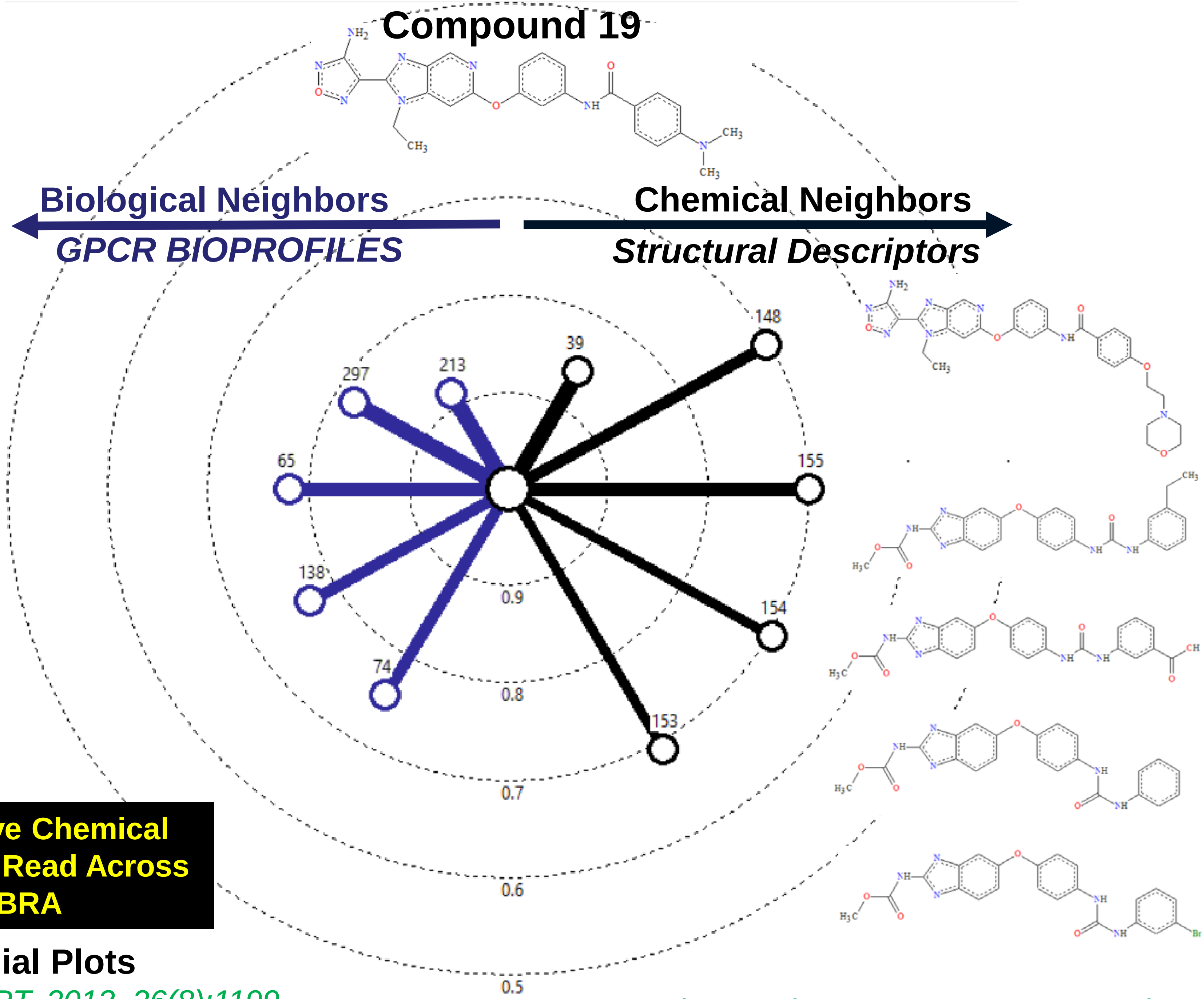
Integrative Chemical Biological Read Across CBRA

CBRA Radial Plots

Low et al. CRT. 2013, 26(8):1199

Fourches et al. JCIM, 2016, In Preparation

Compound 19



Biological Neighbors

GPCR BIOPROFILES

Chemical Neighbors

Structural Descriptors

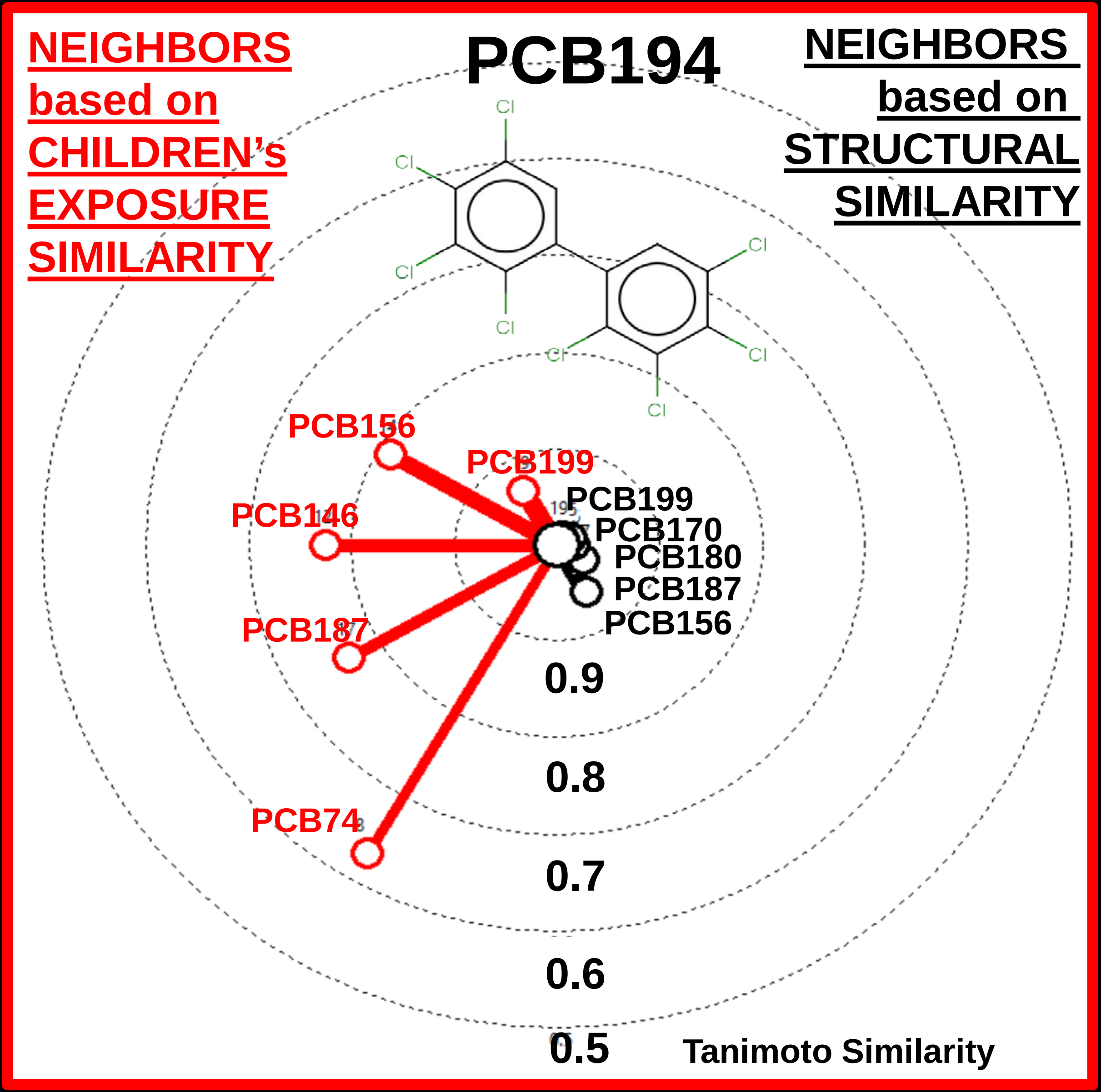
Integrative Chemical Biological Read Across CBRA

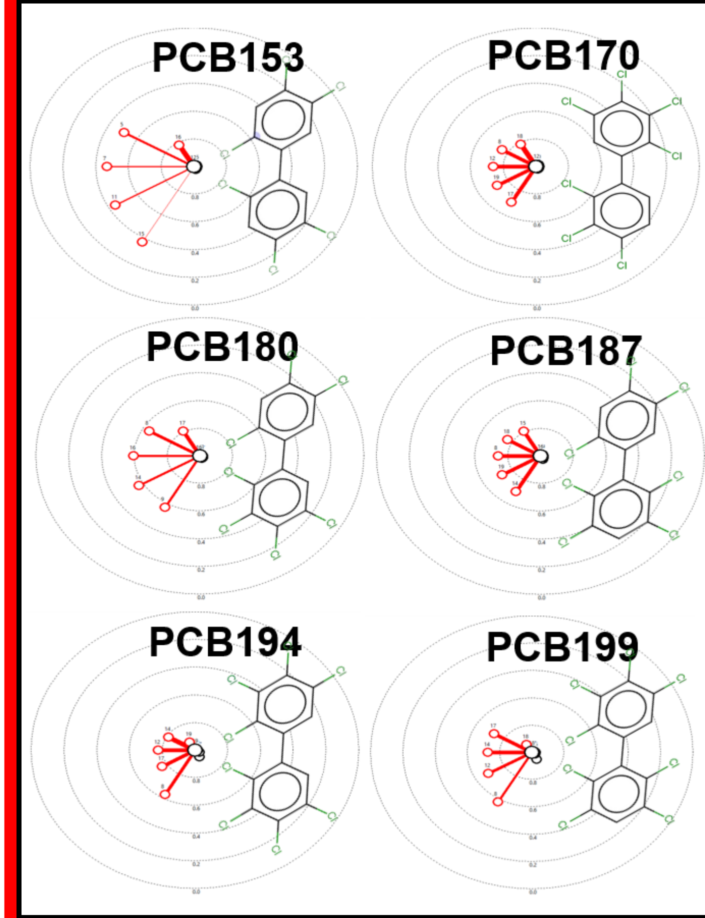
CBRA Radial Plots

Low et al. CRT. 2013, 26(8):1199

Fourches et al. JCIM, 2016, In Preparation

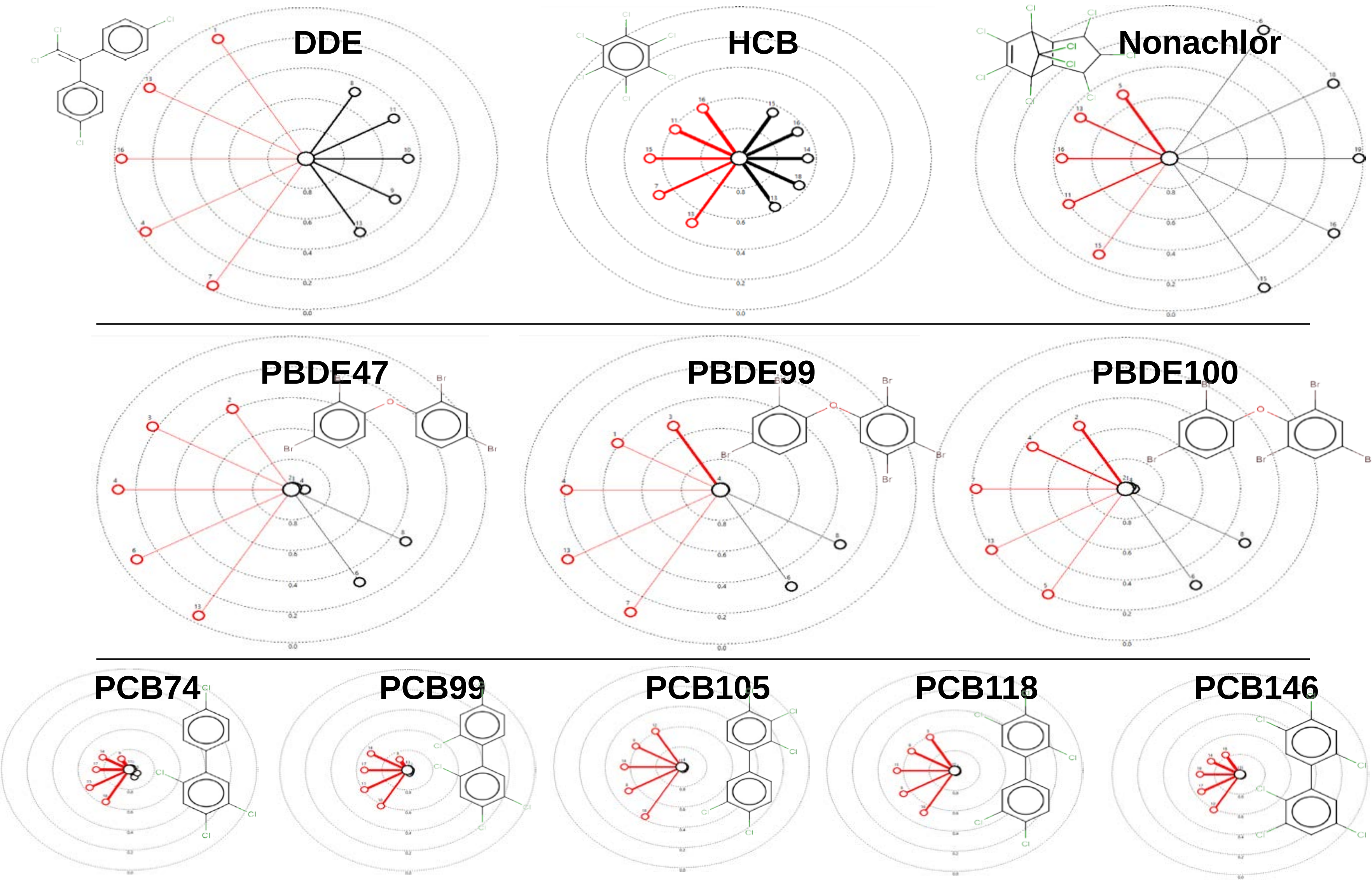
Cheminformatics Approaches To Analyze the Similarity and the Effects of Environmental Chemical Mixtures



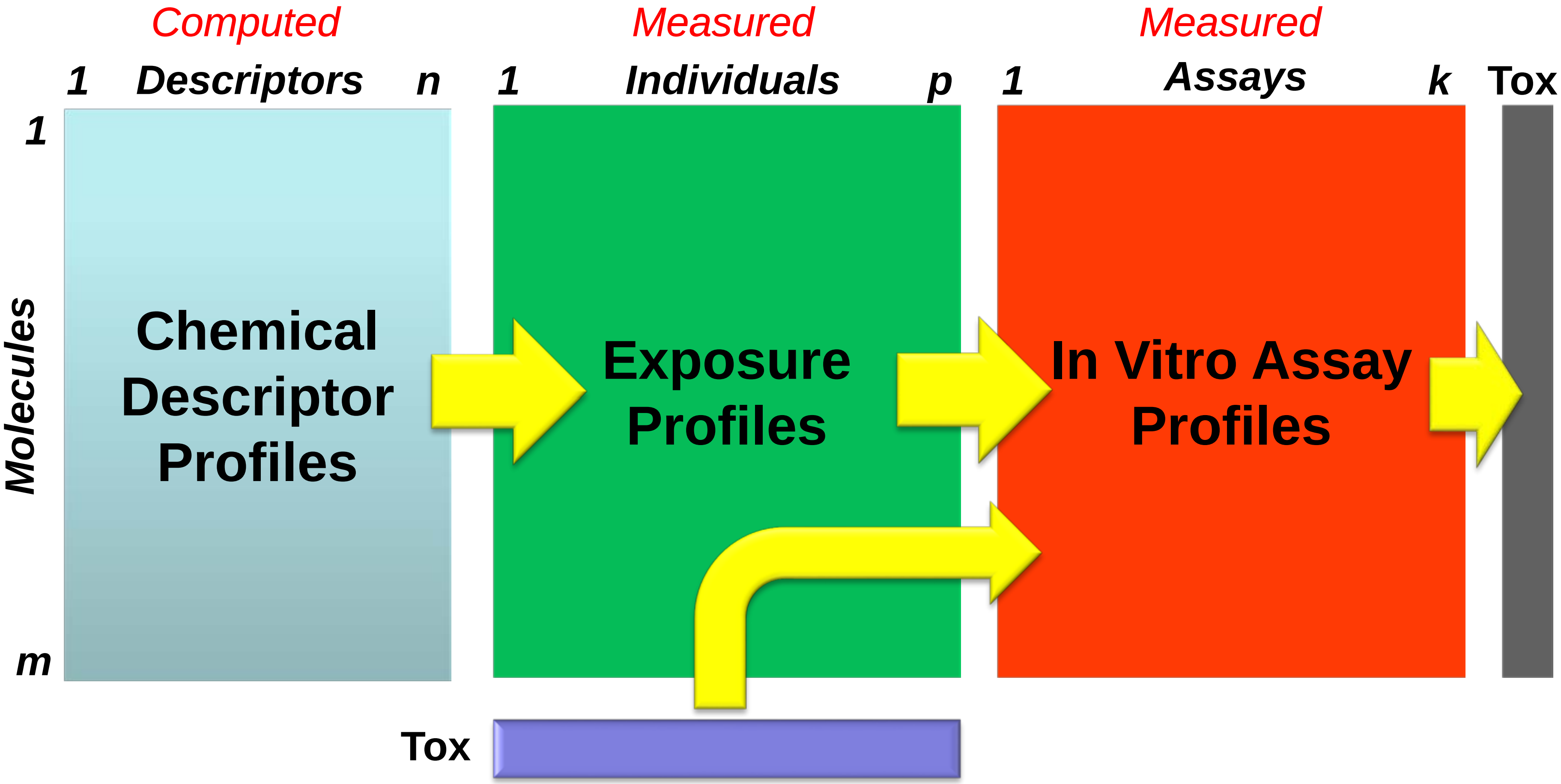


Cheminformatics Approaches To Analyze the Similarity and the Effects of Environmental Chemical Mixtures

CBRA Radial Plots Based on Both Chemical and Children's Exposures Similarity

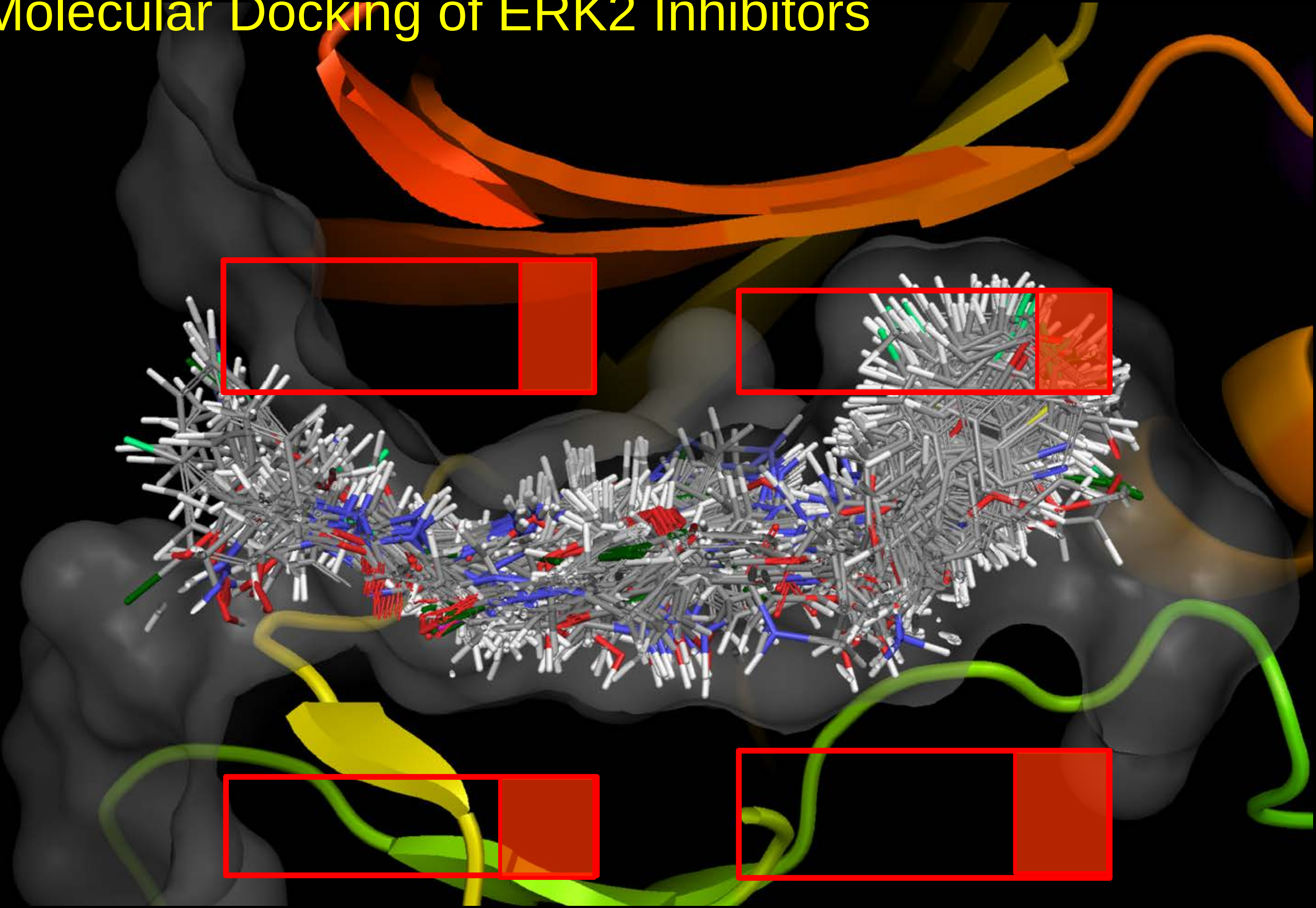


Concept of Quantitative Structure-Exposure-Toxicity Relationships (QSETR)



Need to develop new modeling workflow …

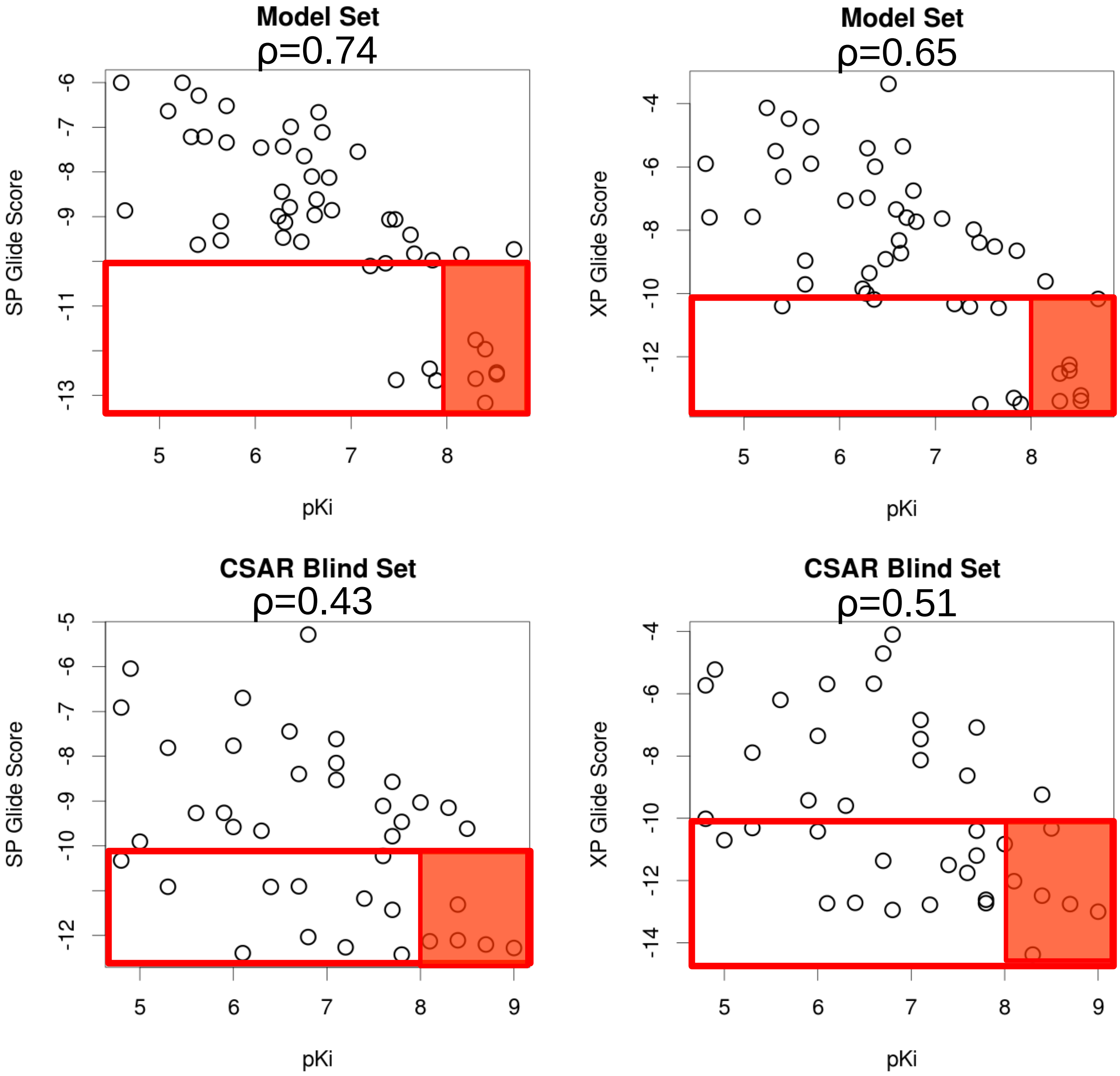
Molecular Docking of ERK2 Inhibitors



Fourches et al. J Chem Inf Model. 2013 Aug 26;53(8):1915-22.

Molec

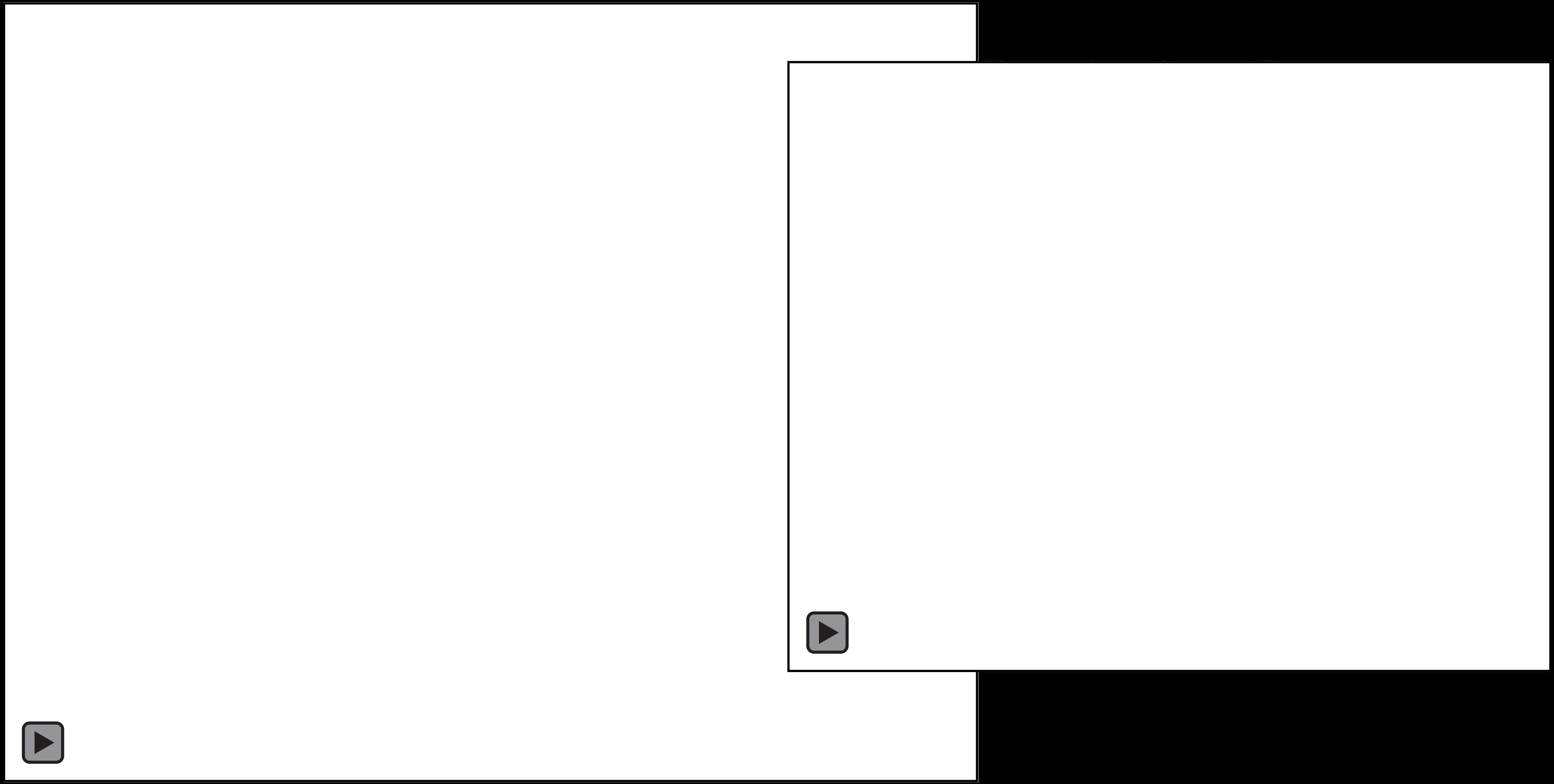
ERK2 – Bad results with Glide 2015-1



Four

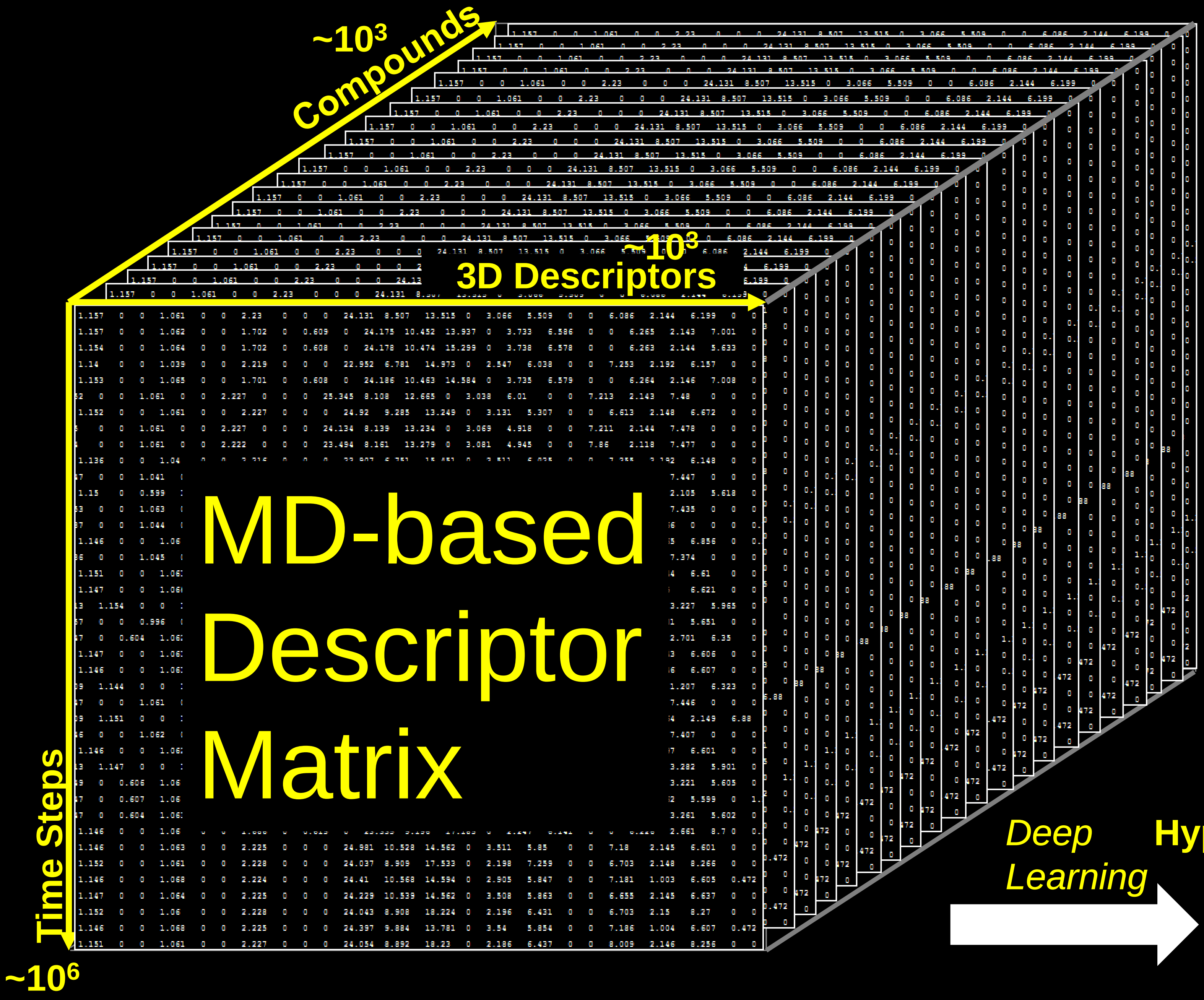
GPU-accelerated molecular dynamics simulations

Example: 1 ns simulation of ERK2 using Desmond
*65Å*90Å*70Å, 42k atoms, explicit solvent (TIP3P water), step= 1 femtosecond*



Up to 1 μs per day on high-end GPU workstations!

New MD-QSAR modeling approach



~10^3 Compounds

~10^3 3D Descriptors

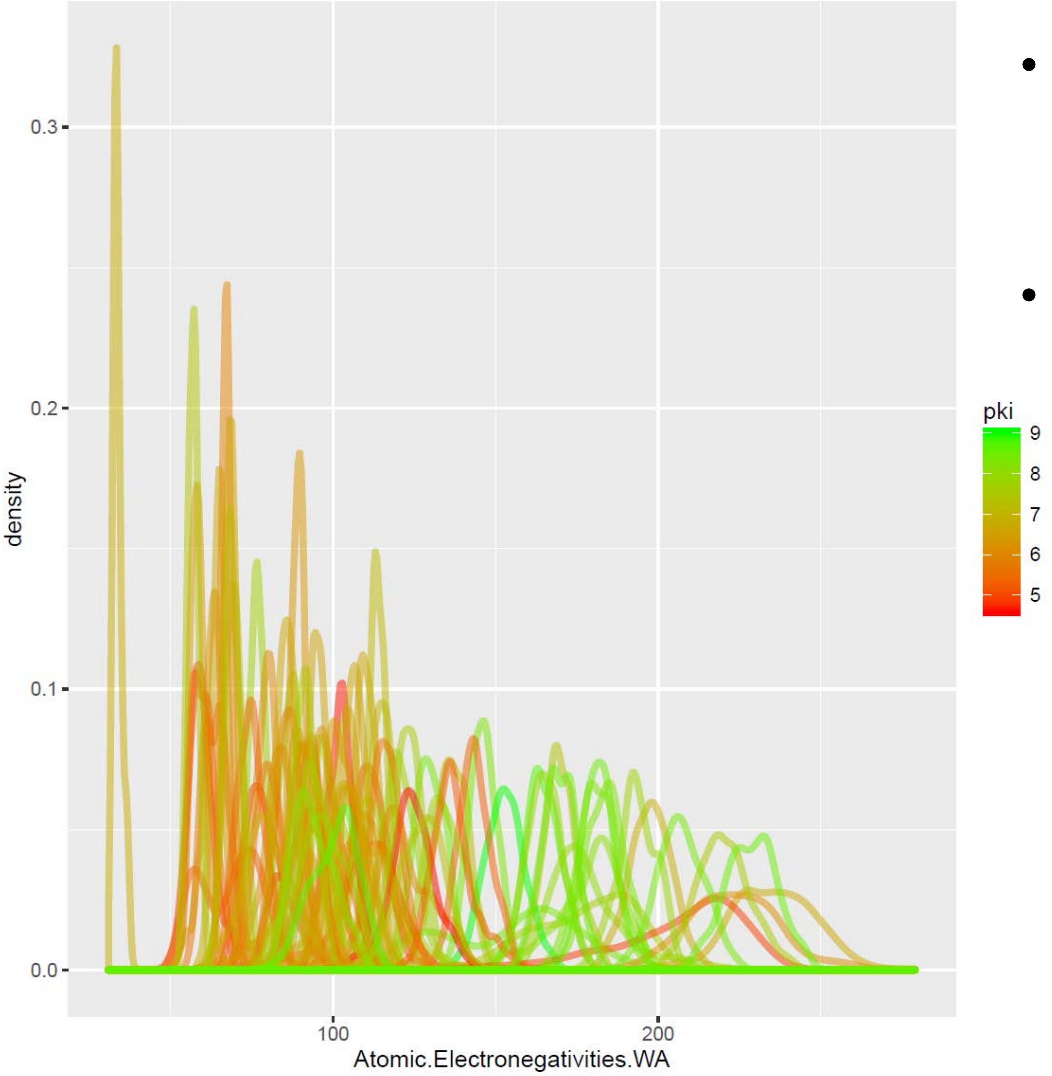
~10^6 Time Steps

MD-based Descriptor Matrix

Deep Learning

Hyper-Predictive QSAR models

Computation of MD Descriptors



- 3D descriptors computed for all conformations sampled along ligand MD simulations

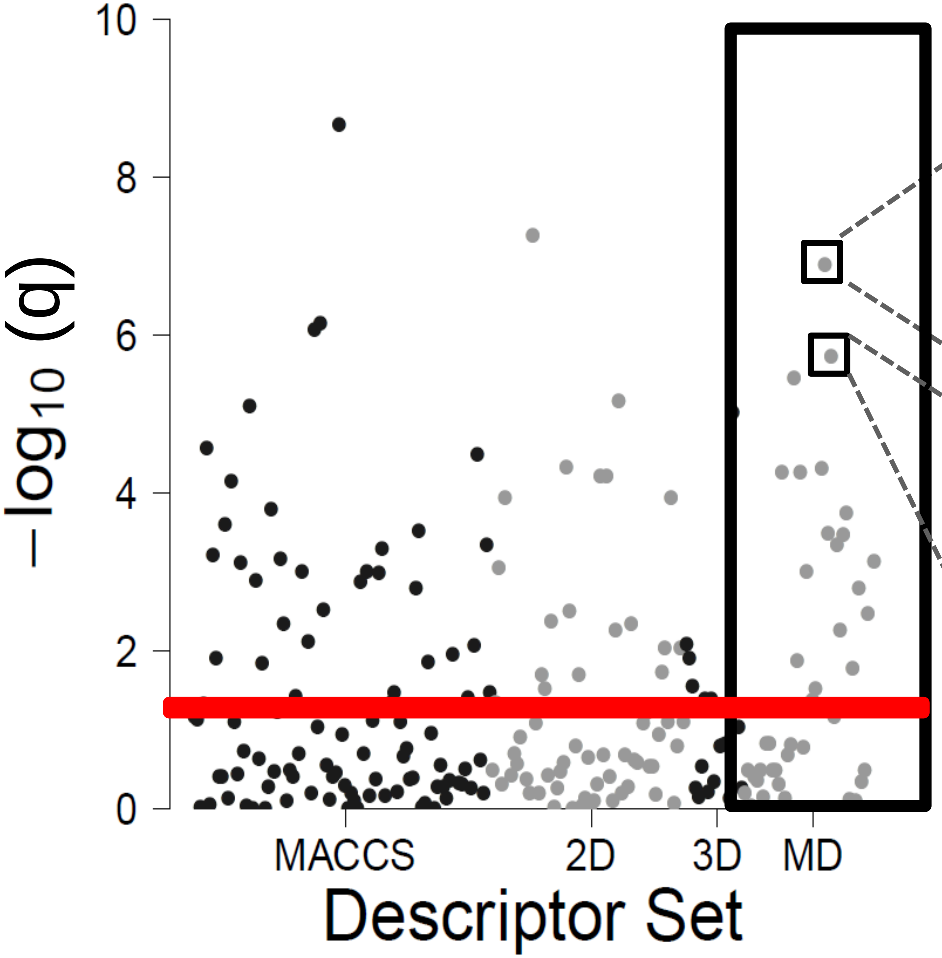
- MD descriptors were constructed by taking the mean and standard deviation of each 3D descriptor distribution for each ligand:

$$\bar{x}_i = \frac{\sum_{j=1}^{n} x_{ij}}{n}$$

$$s_i = \sqrt{\frac{\sum_{j=1}^{n}(x_{ij} - \bar{x})^2}{n-1}}$$

Chemical Descriptors Associated with Ligand Activity

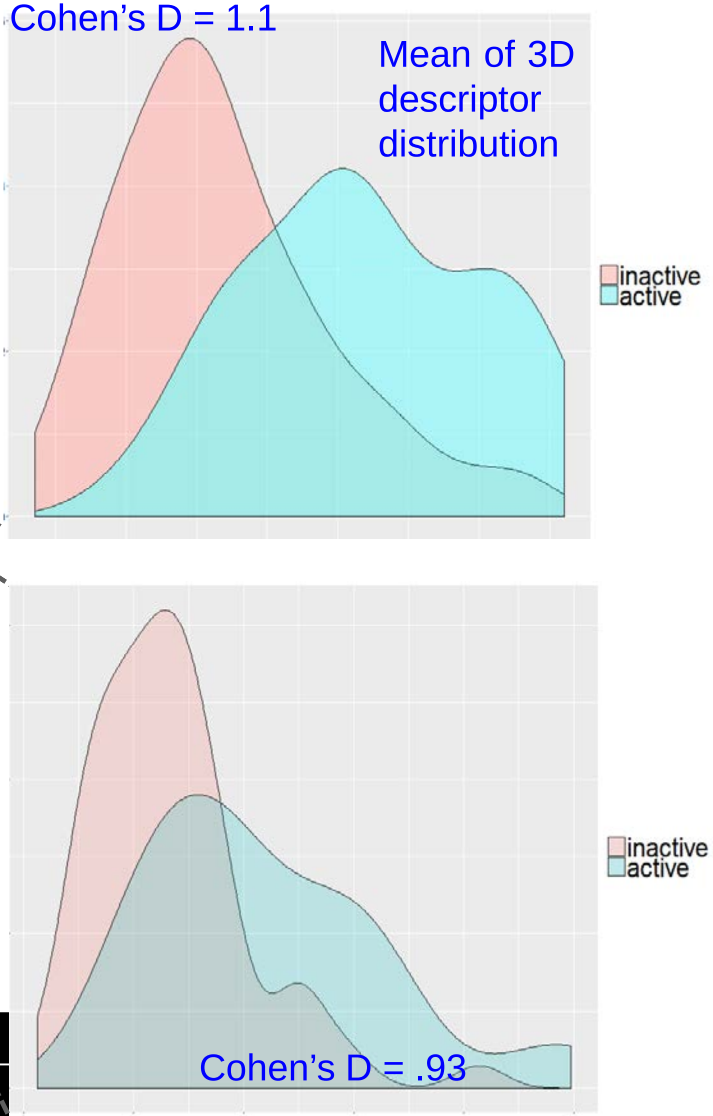
Dataset: 87 ERK2 kinase inhibitors; pKi ranging from 4.6 to 9



Atomic Masses Weighted WV

Cohen's D = 1.1

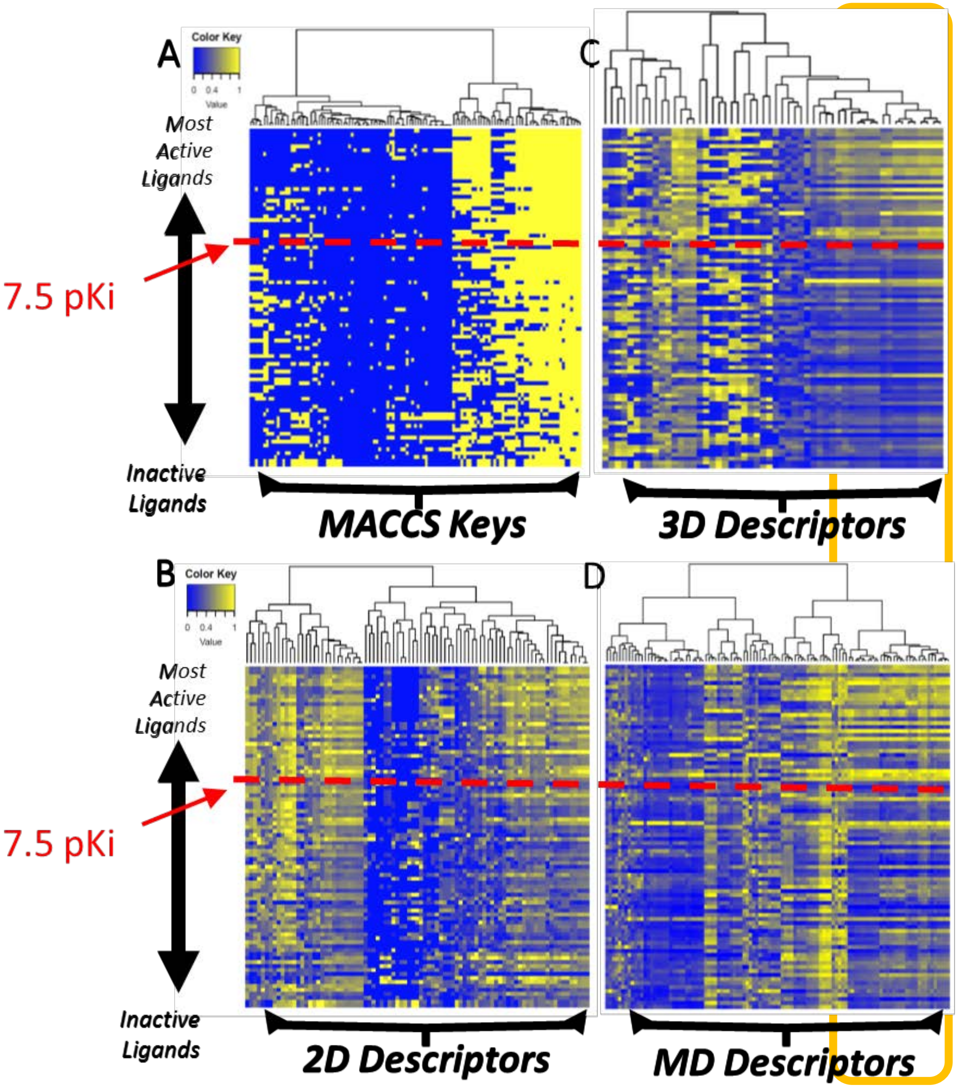
Mean of 3D descriptor distribution



Cohen's D = .93

Ash and Fourches. JCIM, 2017. Under Review

Descriptor Set Distributions

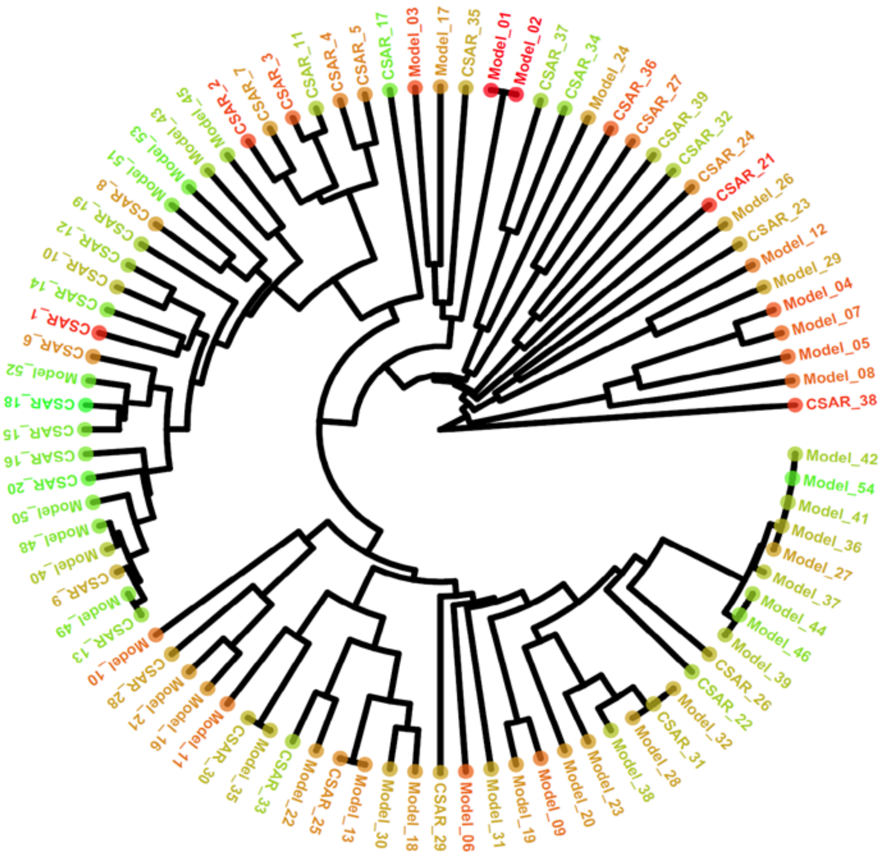


- Compounds were classified as
  - Active : pKi >= 7.5
  - Inactive : pKi < 7.5
- For the 3D and MD descriptors, clear difference in the profiles of active and inactive compounds.
- For MACCS and 2D descriptors, the difference in active and inactive profiles is less apparent

Ash and Fourches. JCIM 2017. Under Review

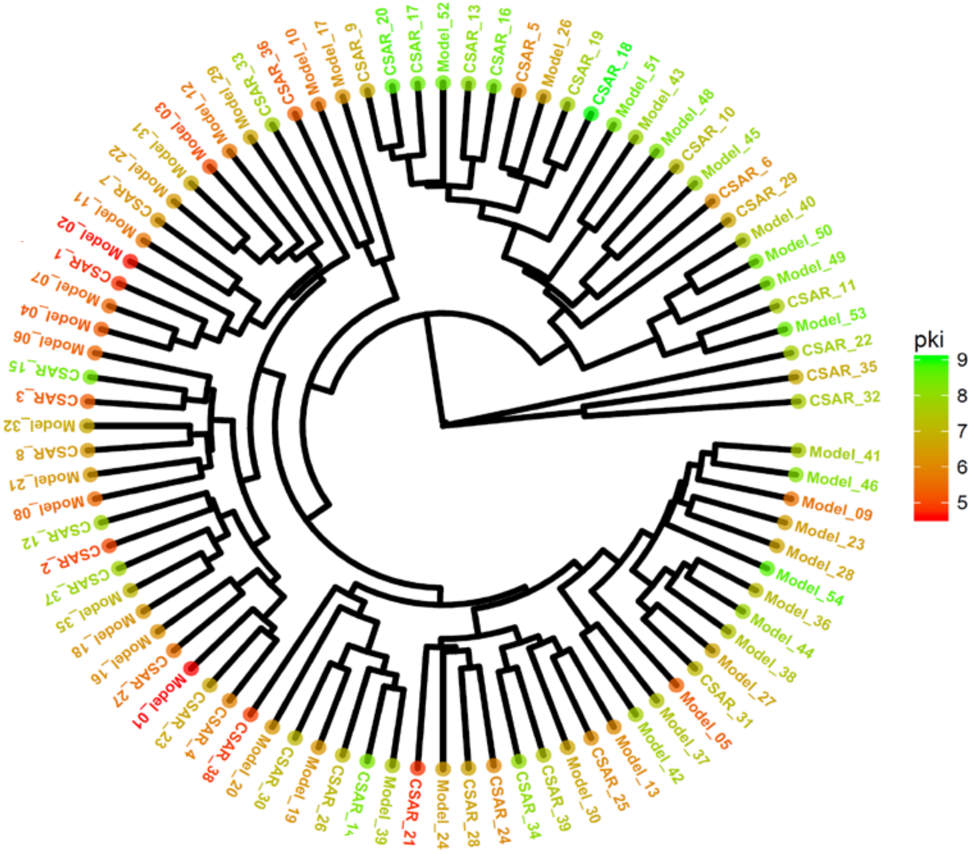
Hierarchical Clustering of ERK2 Ligands Using 2D and MD fingerprints

MACCS Fingerprints



Cophenetic correlation coefficient: **0.89**

MD Descriptors

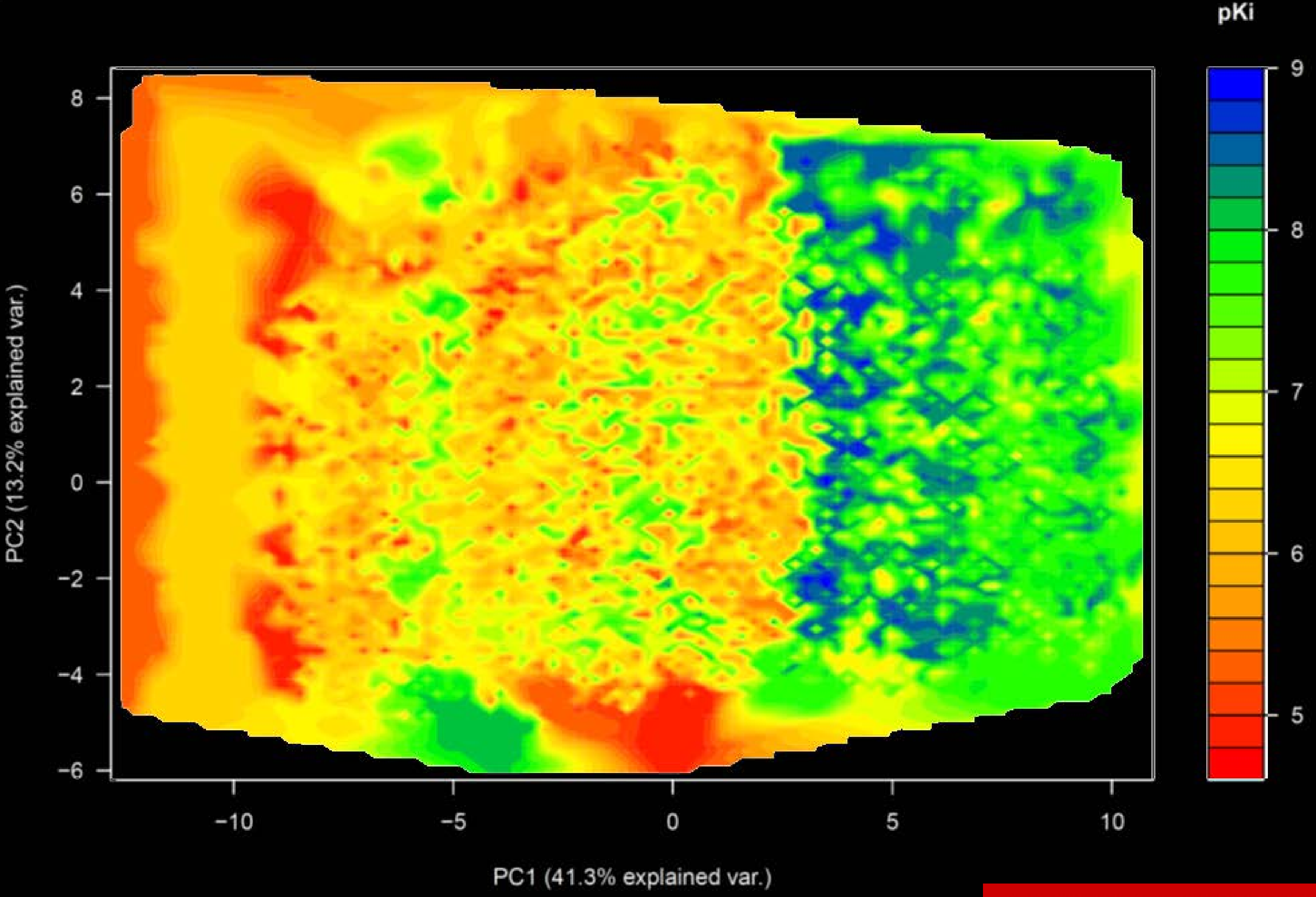


Cophenetic correlation coefficient: **0.74**

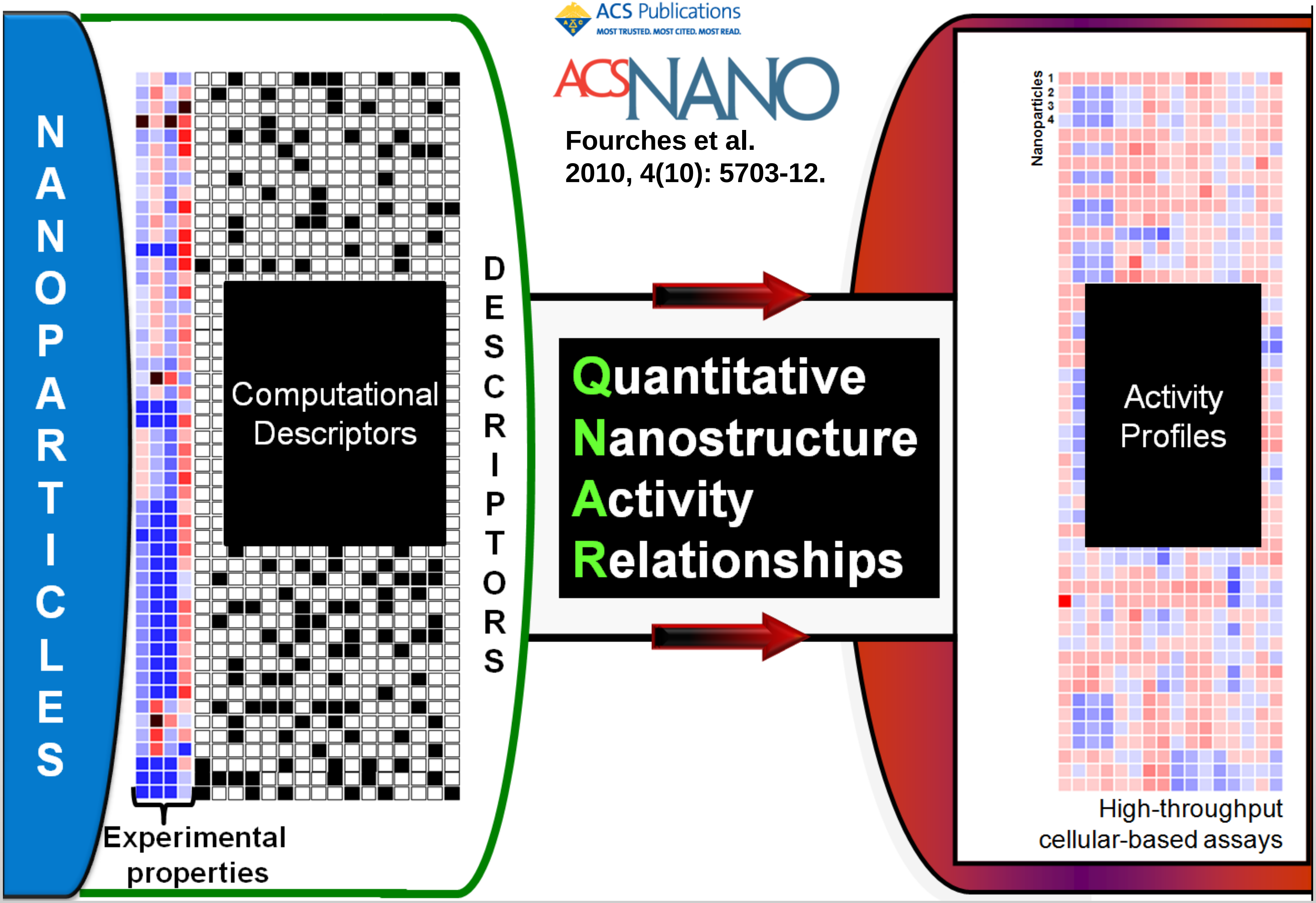
Ash and Fourches. JCIM 2017. Under Review

NC STATE UNIVERSITY

Characterizing the MD Chemical Space of ERK2 Inhibitor Conformations



Ash and Fourches. JCIM 2017. Under Review

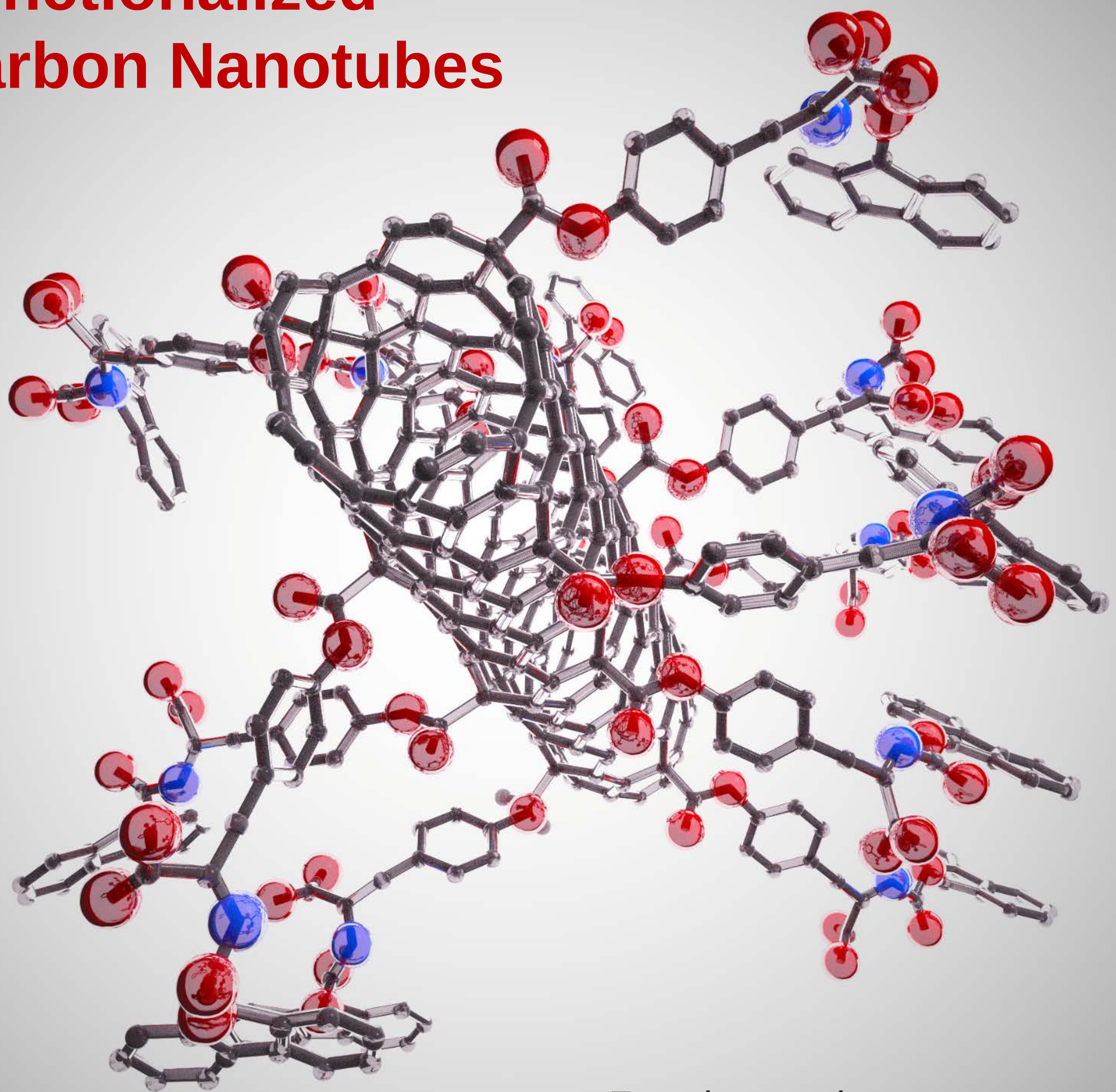


Fourches et al.
2010, 4(10): 5703-12.

Quantitative
Nanostructure
Activity
Relationships

Fourches D, Pu D, Tropsha A. *Comb Chem High Throughput Screen*. 2011
Fourches et al. *Nanotoxicology*, 2015, In Press.

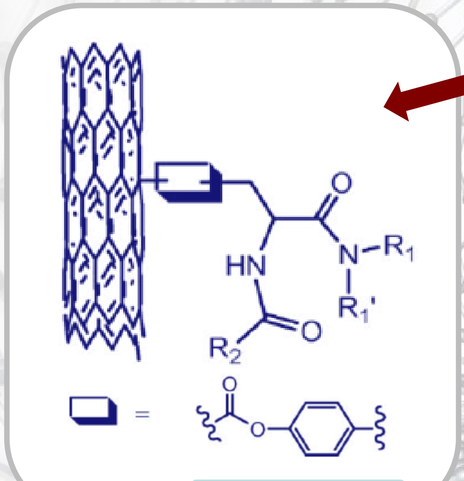
Functionalized Carbon Nanotubes

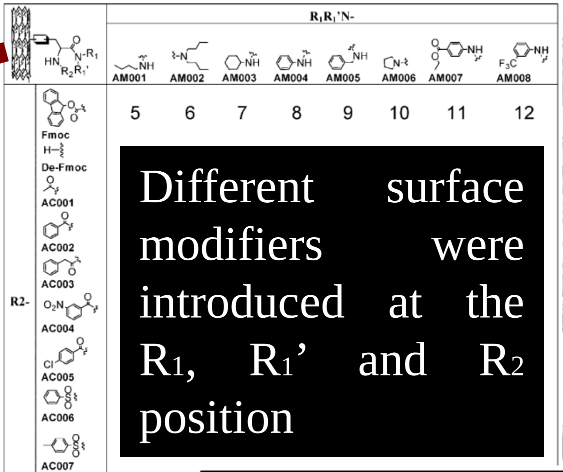


Fourches et al.
Nanotoxicology. In Preparation

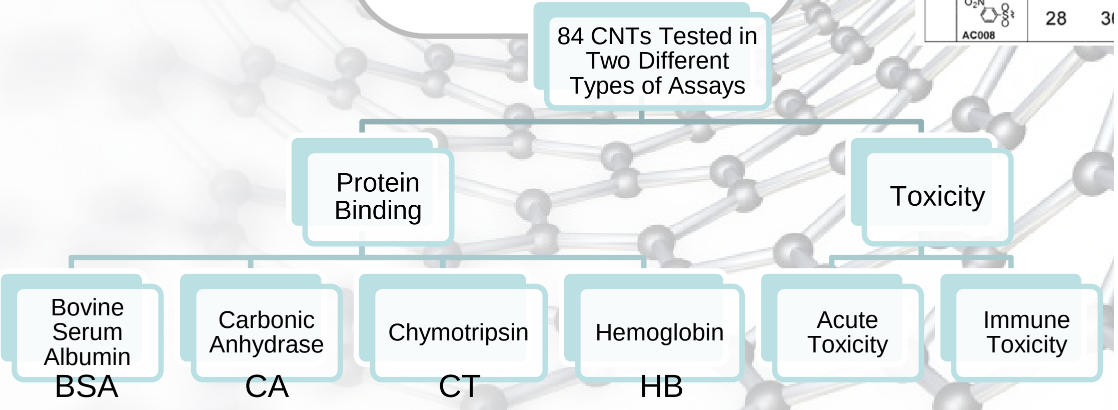
QNAR Modeling of Carbon Nanotubes

In 2008, Zhou et al* published *in vitro* protein binding, acute toxicity and immune toxicity assays for 84 Carbon NanoTubes (CNTs) decorated with different surface modifiers.




R1R1'N-
HN
N-R1
R2R1'
AM001
AM002
AM003
AM004
AM005
AM006
AM007
AM008
R2-
Fmoc
De-Fmoc
AC001
AC002
AC003
AC004
AC005
AC006
AC007
AC008
5
6
7
8
9
10
11
12
28
30

Different surface modifiers were introduced at the R_1, R_1' and R_2 position



84 CNTs Tested in Two Different Types of Assays

Protein Binding

Toxicity

Bovine Serum Albumin BSA

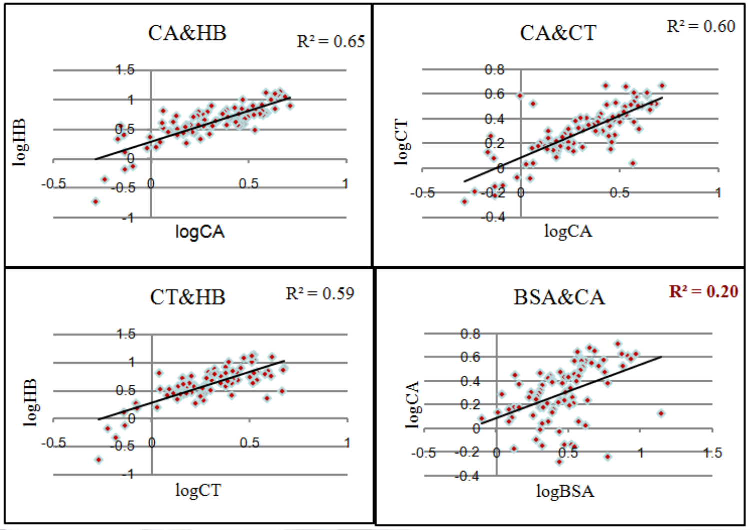
Carbonic Anhydrase CA

Chymotripsin CT

Hemoglobin HB

Acute Toxicity

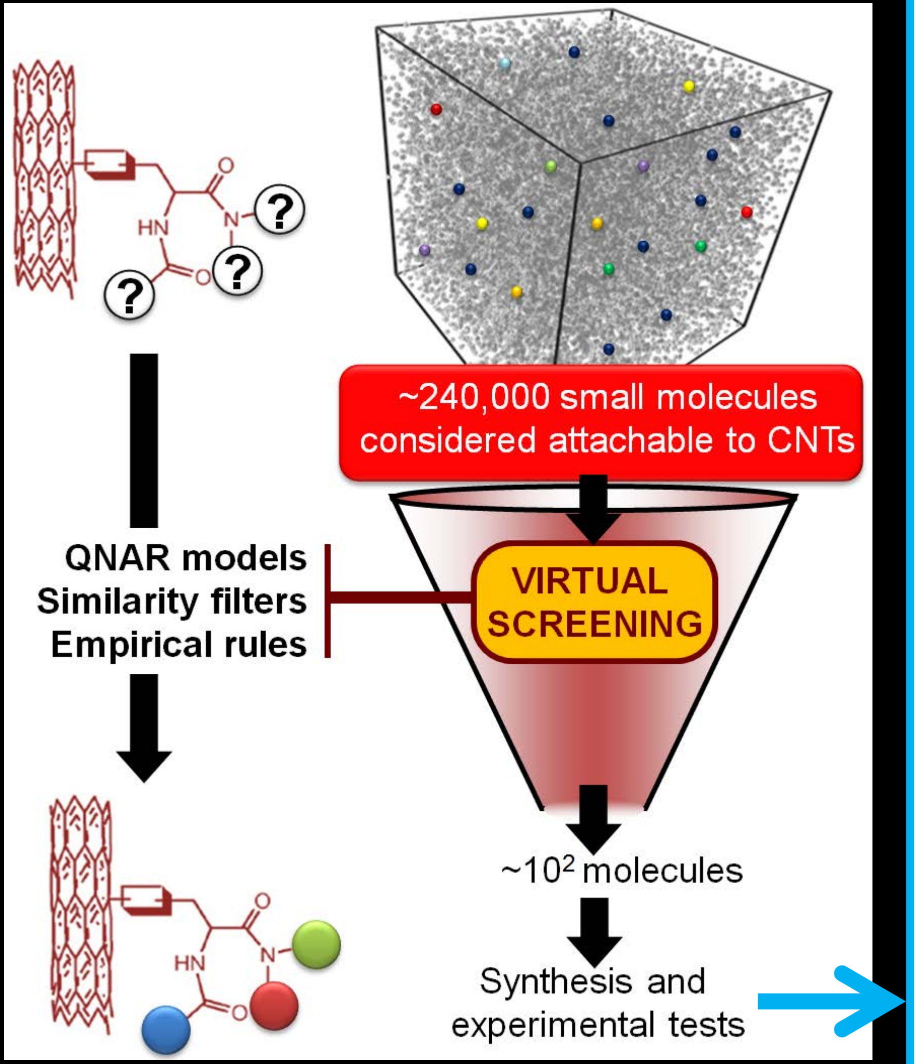
Immune Toxicity

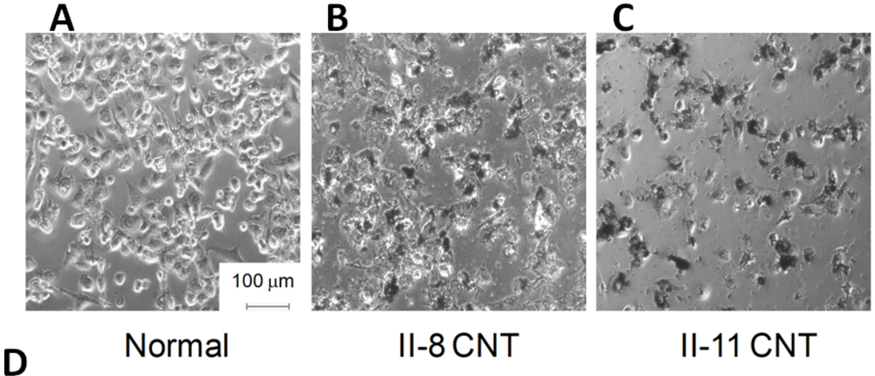


Zhou et al. Nano Lett., Vol. 8, No. 3, 2008

Computer-aided design of novel carbon nanotubes with desired biological properties

(in collaboration with Dr. Bing Yan, St. Jude Children's Research Hospital)





D

CNT ID	II-1	II-2	II-3	II-4	II-5	II-6	II-7	II-8	II-9	II-10
Average cell viability (%)	58	61	61	56	58	65	68	72	68	59
Standard Deviation (%)	5	3	3	3	2	10	6	7	3	6
Experiment	0	0	0	0	0	0	0	0	0	0
Predicted	0	0	0	0	0	0	0	0	0	0
CNT ID	II-11	II-12	II-13	II-14	II-15	II-16	II-17	II-18	II-19	II-20
Average cell viability (%)	39	49	46	49	52	41	51	49	55	50
Standard Deviation (%)	9	8	7	5	8	11	5	9	11	10
Experiment	1	1	1	1	0	1	0	1	0	0
Predicted	1	1	1	1	1	1	1	1	1	1

CNT ID	II-21	II-22	II-23	II-24	II-25	II-26	II-27	II-28	II-29	II-30
Average protein binding (F0/F1)	1.77	1.78	1.87	1.76	1.82	2.30	1.74	2.00	1.67	2.33
Standard Deviation	0.05	0.06	0.02	0.03	0.02	0.02	0.02	0.01	0.06	0.02
Experiment	0	0	0	0	0	1	0	1	0	1
Predicted	0	0	0	0	0	0	0	0	0	0
CNT ID	II-31	II-32	II-33	II-34	II-35	II-36	II-37	II-38	II-39	II-9
Average protein binding (F0/F1)	3.40	2.28	2.04	2.22	2.17	2.95	2.08	2.25	2.65	2.24
Standard Deviation	0.03	0.05	0.08	0.01	0.04	0.00	0.02	0.06	0.05	0.11
Experiment	1	1	1	1	1	1	1	1	1	1
Predicted	1	1	1	1	1	1	1	1	1	1

Fourches et al. Nanotoxicology 2016 In Press.

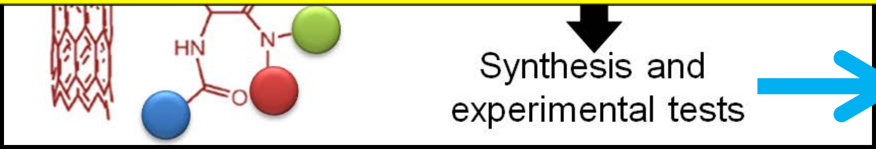
Computer-aided design of novel carbon nanotubes with desired biological properties

(in collaboration with Dr. Bing Yan, St. Jude Children's Research Hospital)

All rationally prioritized, synthesized, and tested CNTs predicted as non-toxic were confirmed experimentally.

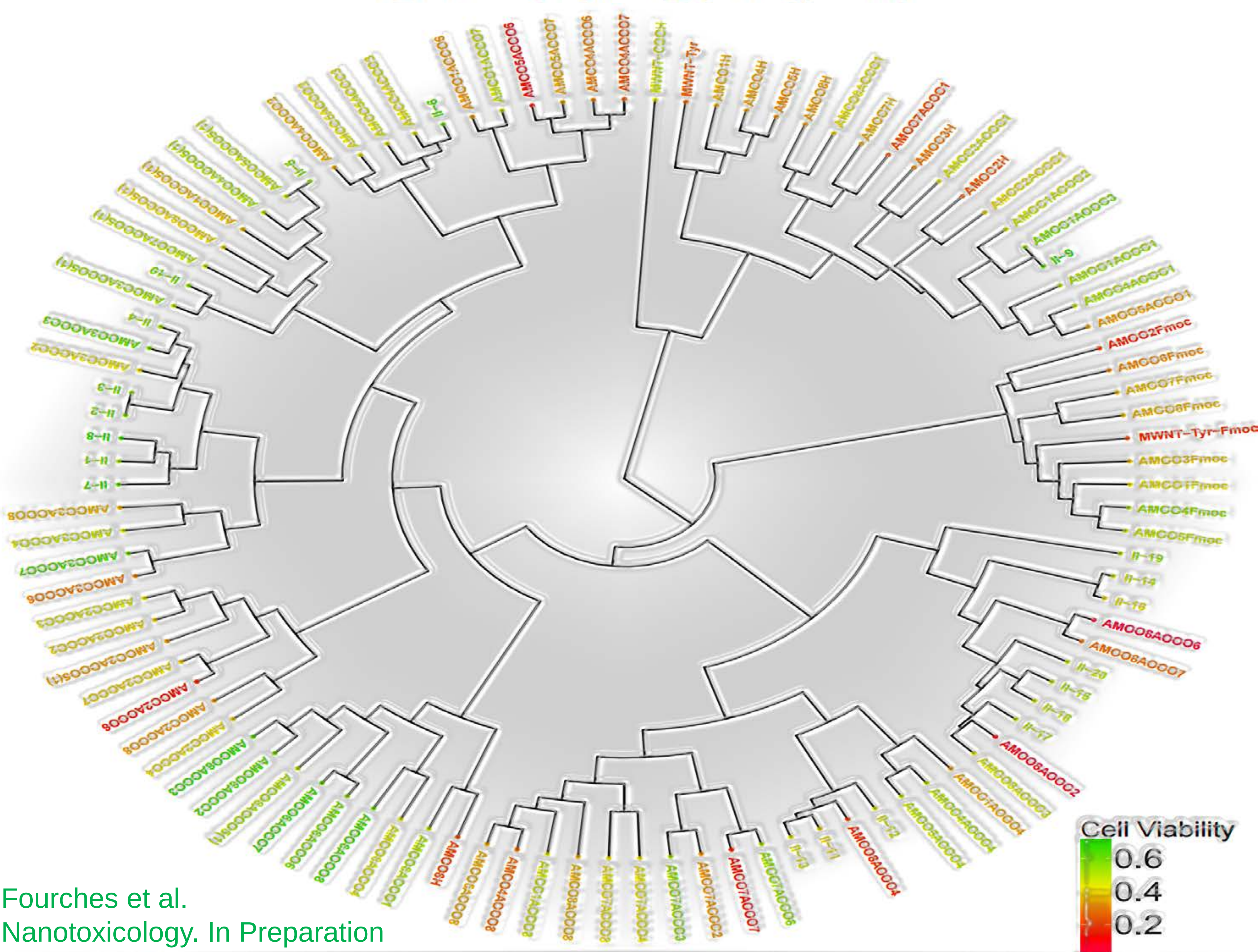
Standard Deviation (%)	5	3	3	3	2	10	6	7	3	6
Experiment	0	0	0	0	0	0	0	0	0	0

6 out of 10 rationally prioritized, synthesized, and tested CNTs predicted as toxic were confirmed experimentally



Experiment	0	0	0	0	0	1	0	1	0	1
Predicted	0	0	0	0	0	0	0	0	0	0
CNT ID	II-31	II-32	II-33	II-34	II-35	II-36	II-37	II-38	II-39	II-9
Average protein binding (F0/F1)	3.40	2.28	2.04	2.22	2.17	2.95	2.08	2.25	2.65	2.24
Standard Deviation	0.03	0.05	0.08	0.01	0.04	0.00	0.02	0.06	0.05	0.11
Experiment	1	1	1	1	1	1	1	1	1	1
Predicted	1	1	1	1	1	1	1	1	1	1

Fourches et al. Nanotoxicology 2016 In Press.



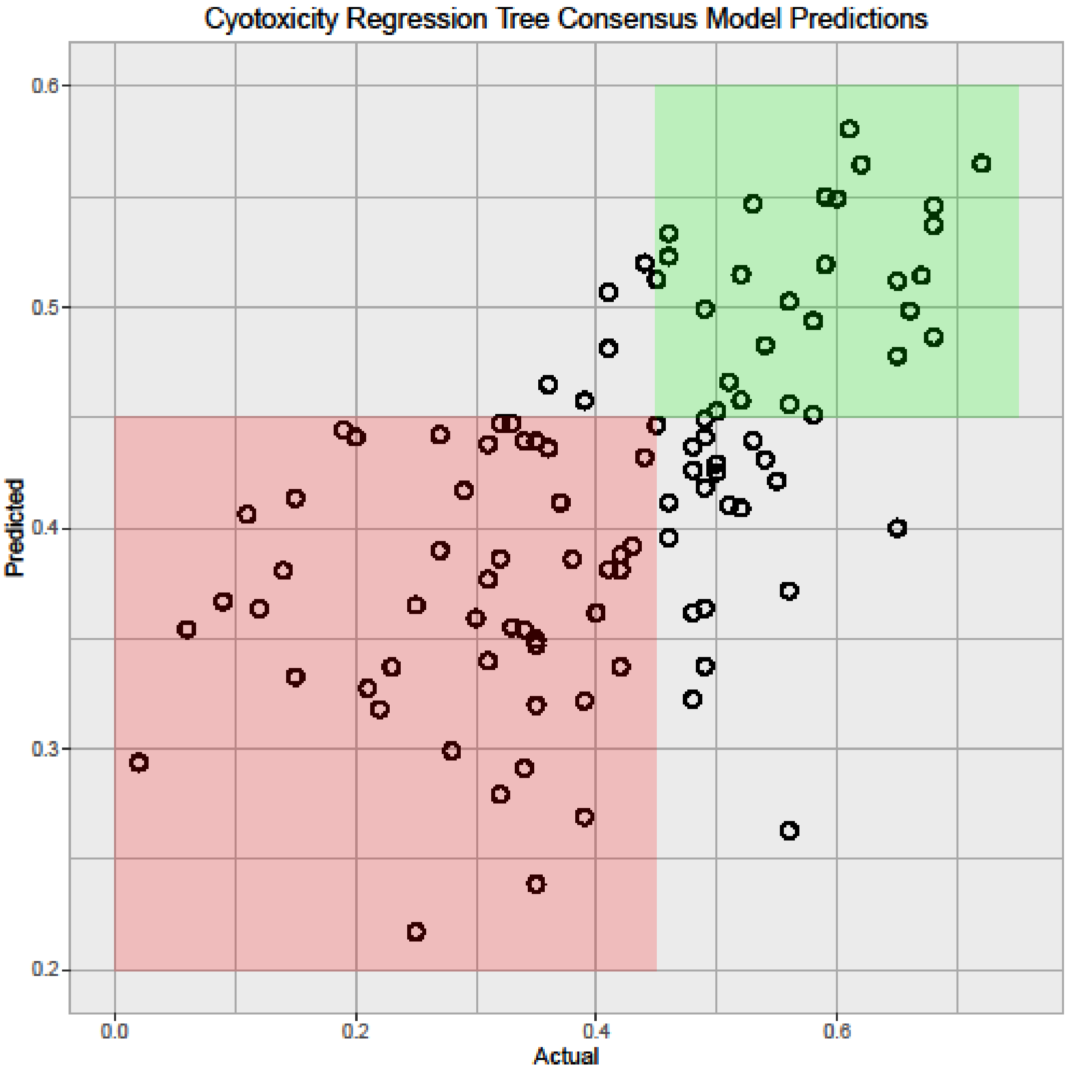
Fourches et al.
Nanotoxicology. In Preparation

QNAR models for the enhanced set of f-CNTs

Classification Models (LOO)

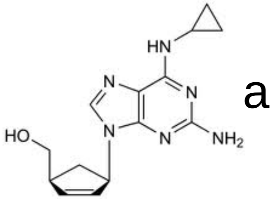
Cytotoxicity	Threshold	Sensitivity		Specificity		Accuracy
Random Forest	0.45	79.25%	42	62.00%	31	70.87%
SVM-C Model	0.45	74.40%	41	56.00%	28	66.99%

Continuous Models (LOO)

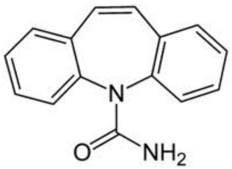


HLA-induced drug adverse effects

Small molecule drugs bind specifically to certain type of HLA proteins.

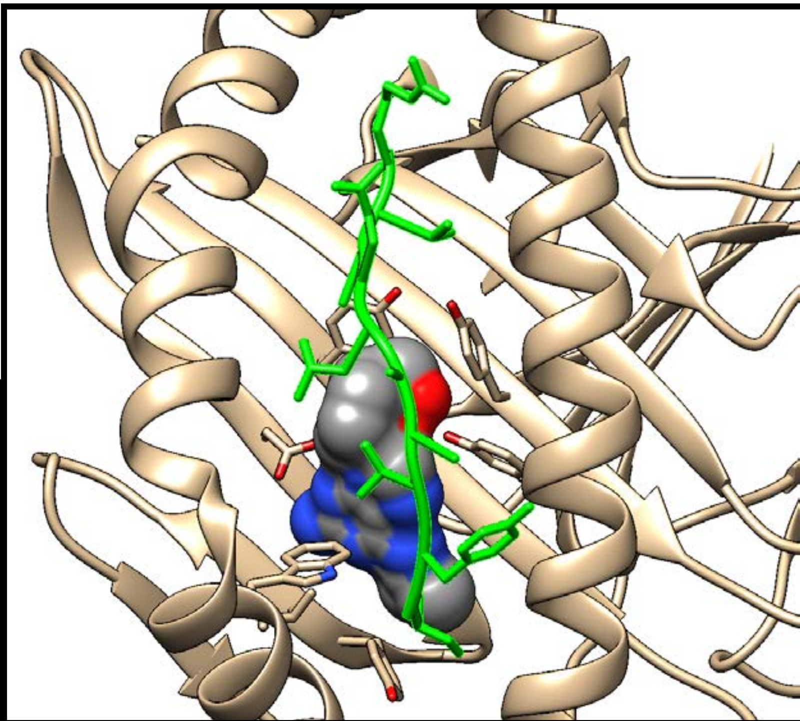


abacavir → HLA-B*57:01 → Abacavir Hypersensitivity Syndrome (AHS)



carbamazepine → HLA-B*15:02 → Stevens-Johnson Syndrome (SJS)

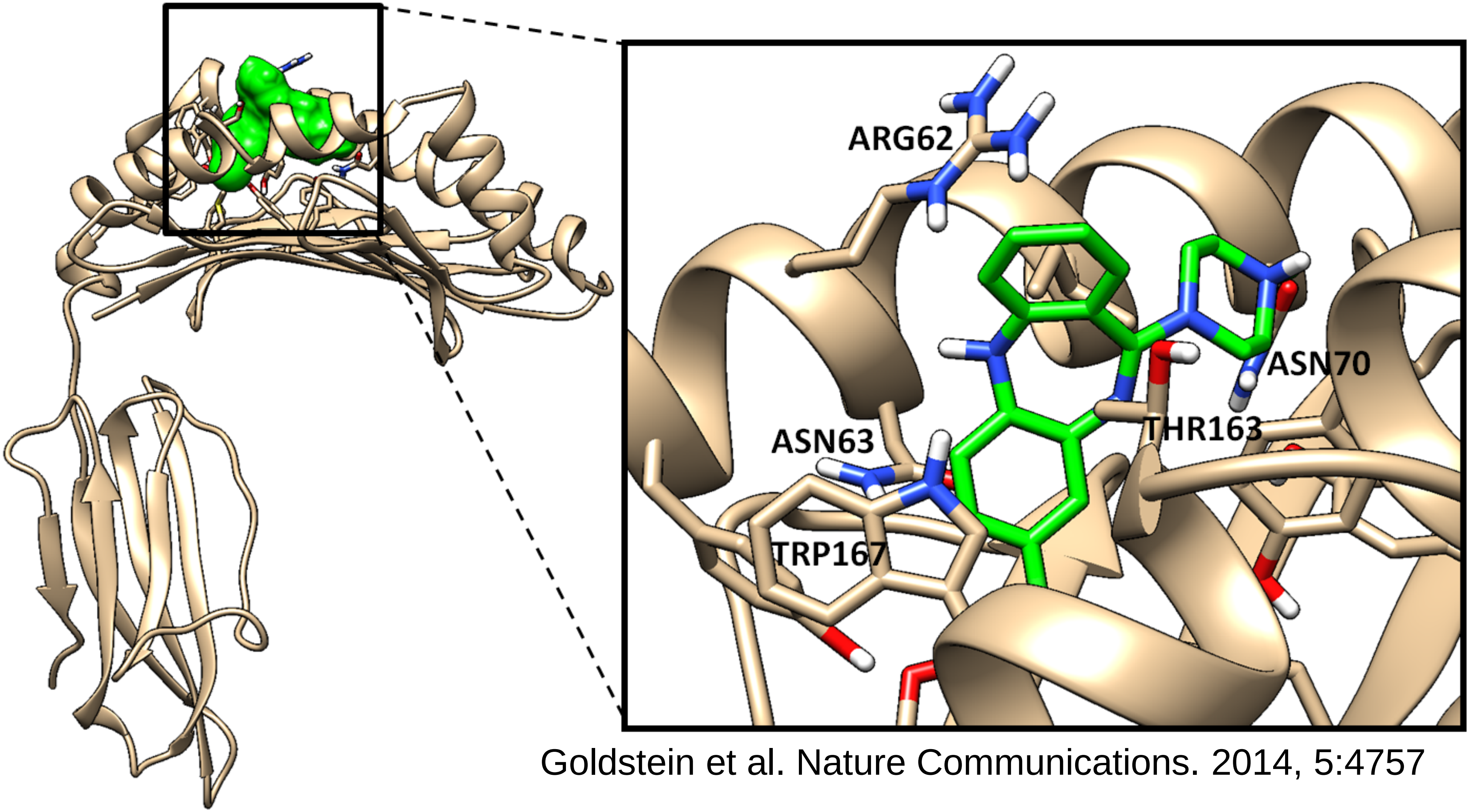
- Binding pocket of HLA proteins is slightly modified
- Range of “self” peptides able to bind the given HLA type is modified
- Immune reaction



- **HLA proteins as important off-targets**
- **Need predictive models to assess HLA-induced ADR**

Clozapine-induced agranulocytosis is associated with rare HLA-DQB1 and HLA-B alleles.

GWAS + Docking

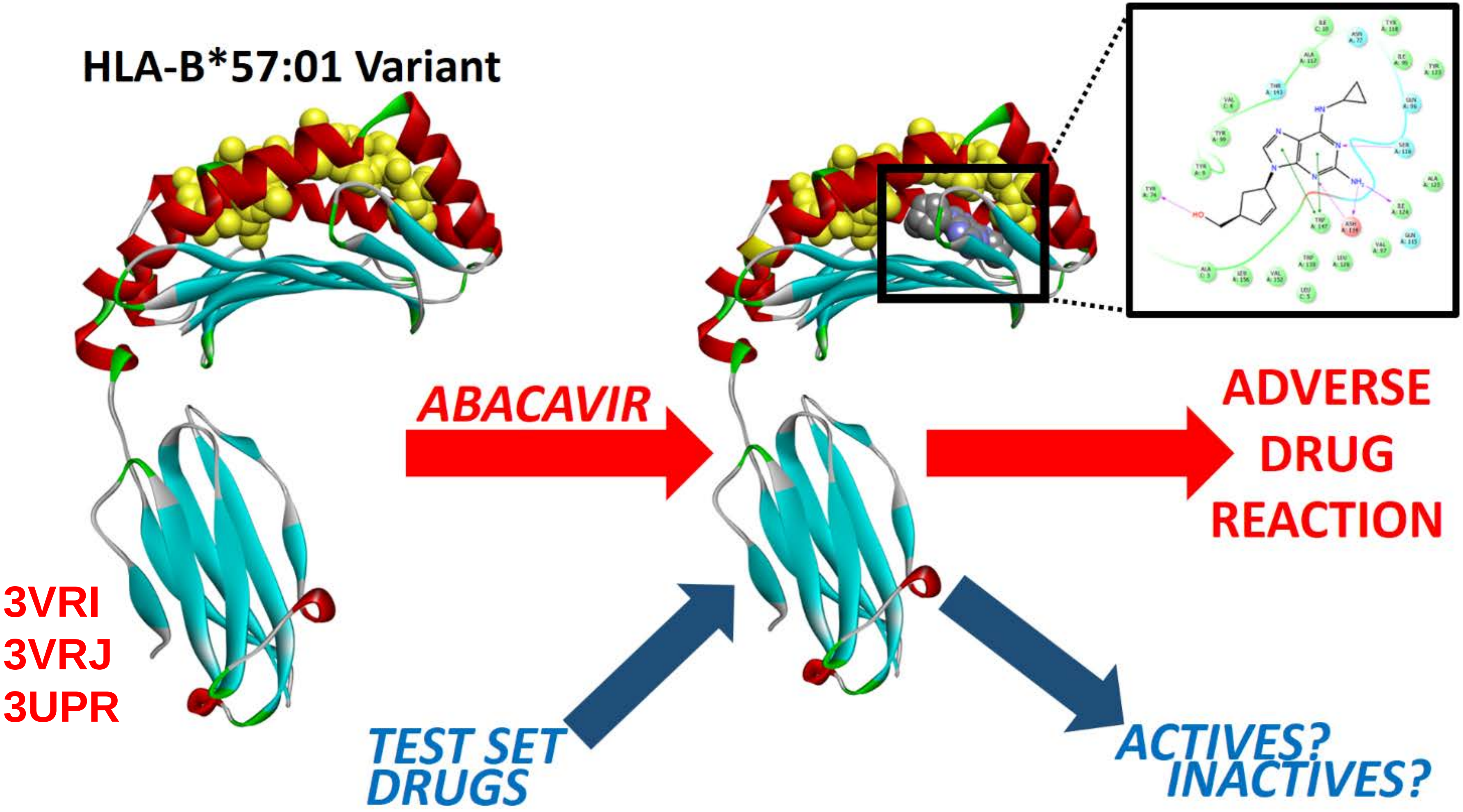


Goldstein et al. Nature Communications. 2014, 5:4757

Homology model of HLA-B39

Molecular Docking Study at HLA-B*57:01

Develop a virtual screening model using molecular docking at the HLA-B*57:01 variant using three X-ray crystals of abacavir bound to HLA-B*57:01.



HLA-B*57:01 Variant

ABACAVIR

ADVERSE DRUG REACTION

3VRI
3VRJ
3UPR

TEST SET DRUGS

ACTIVES?
INACTIVES?

Van Den Driessche et al. 2016, J. Cheminformatics. Under Review.

Summary

- Cheminformatics is becoming mandatory in all projects involving the curation, integration, characterization, analysis, testing, modeling, visualization, screening of chemicals.

- The skyrocketing amount of freely-available data in the public domain is boosting the development of new cheminformatics approaches to fully exploit that data, especially when it comes to chemical risk assessment and *in silico* toxicity predictions.

- New methods such as MD-QSAR or QSETR are poised to boost the prediction performances of cheminformatics predictors.

- Structure-based docking and pan-target screening have never been so relevant for chemical risk assessment, especially for key targets such as ER, AhR, or HLA.

Acknowledgements

Lab members

George Van Den Driessche
Jeremy Ash
Melaine Kuenemann, PhD
Ryan Lougee
Phyo Phyo Kyaw Zin
Bethany Cook

Funding

NCSU CFEP Program
RISF
CMI

NC STATE UNIVERSITY

dfourch@ncsu.edu

www.fourches-laboratory.com

Chancellor’s Faculty Excellence Program