EPA
United States
Environmental Protection
Agency

Recent Cheminformatics development at NCCT applied to ER, AR and physicochemical properties of chemicals

Kamel Mansouri

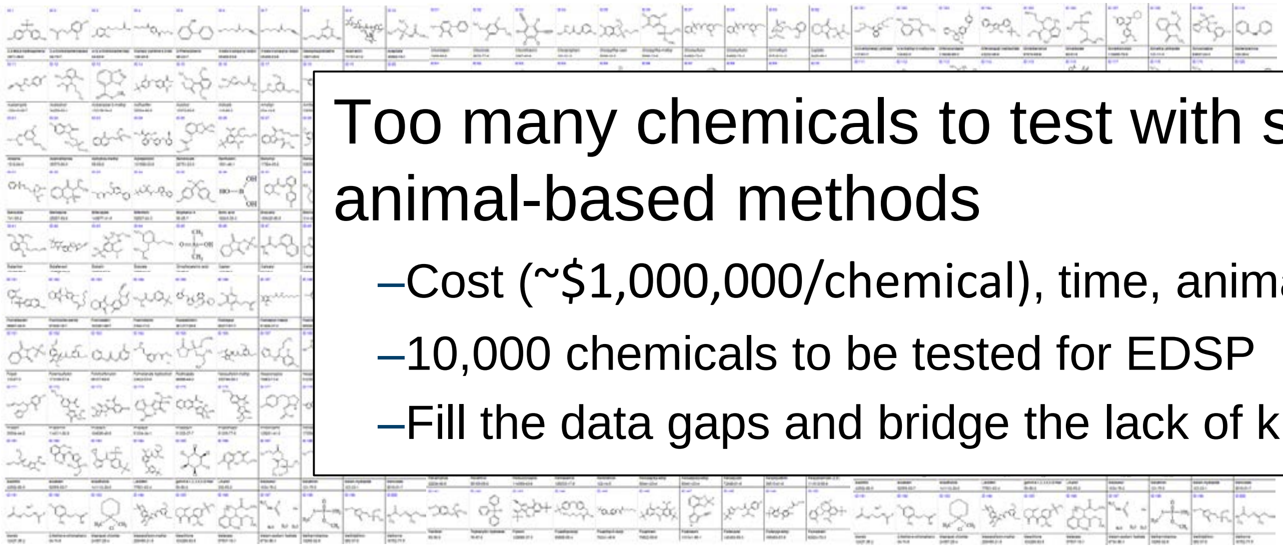
National Center for Computational Toxicology, U.S. EPA, RTP, NC, USA

ORCID ID
orcid.org/0000-0002-6426-8036

Office of Research and Development

The views expressed in this presentation are those of the author and do not necessarily reflect the views or policies of the U.S. EPA

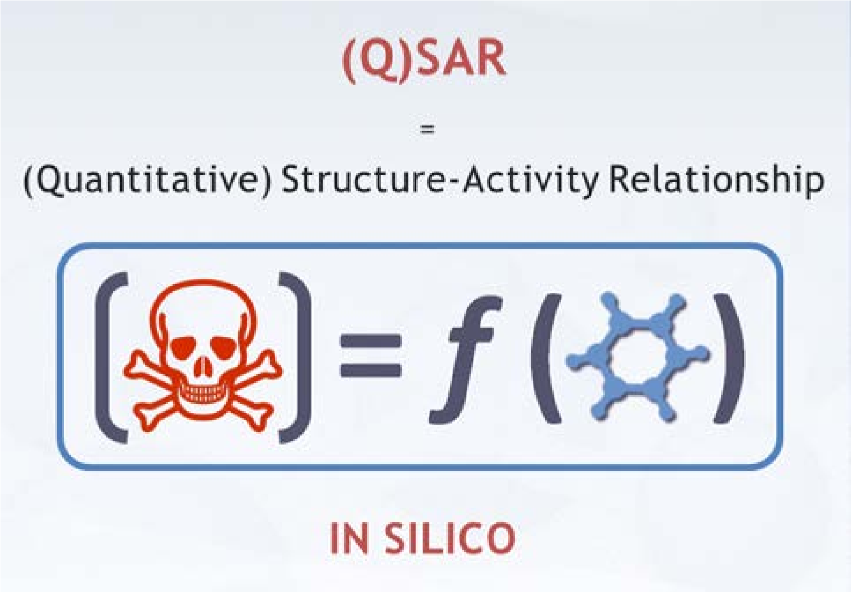
Problem Statement



Too many chemicals to test with standard animal-based methods

- –Cost (~$1,000,000/chemical), time, animal welfare
- –10,000 chemicals to be tested for EDSP
- –Fill the data gaps and bridge the lack of knowledge

Alternative



United States
Environmental Protection
Agency

Recent Cheminformatics development at NCCT

- We are building a new cheminformatics architecture
- PUBLIC dashboard gives access to curated chemistry
- Focus on integrating EPA *and* external resources
- Aggregating and curating data, visualization elements and “services” to underpin other efforts
  - RapidTox
  - Read-across
  - Predictive modeling
  - Non-targeted screening

Quantitative Structure Activity/Property Relationships (QSAR/QSPR)

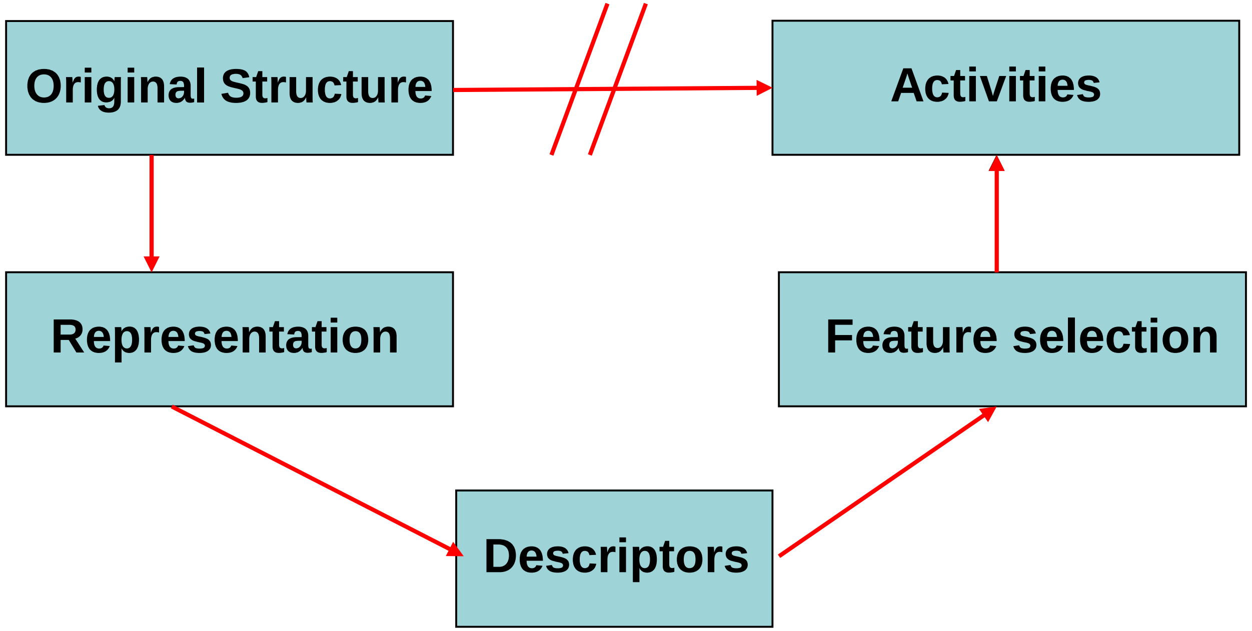
Congenericity principle: QSARs correlate, within congeneric series of compounds, their chemical or biological activities, either with certain structural features or with atomic, group or molecular descriptors.

Katritzky, A. R.; Lobanov, V. S.; Karelson, M. Chem. Soc. Rev. ***1995****, 279-287*

$$Y = f(b_i\,,\,X\,)$$

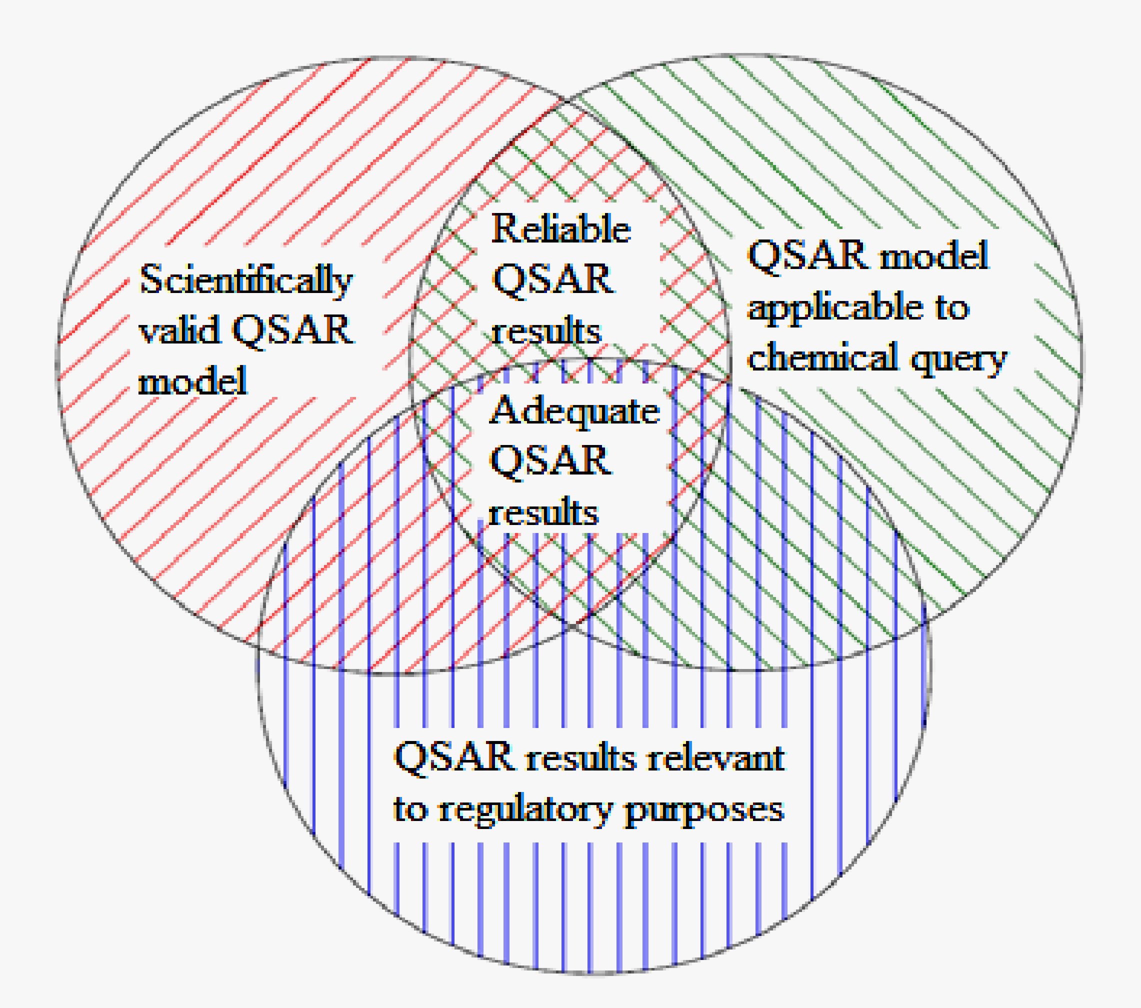
***X - descriptors** (selected variables)*

b_i - fitted parameters



QSARs validity, reliability, applicability and adequacy for regulatory purposes

ORCHESTRA. Theory, guidance and application on QSAR and REACH; 2012. http://home.deib.polimi.it/gini/papers/orchestra.pdf.


Scientifically valid QSAR model
Reliable QSAR results
QSAR model applicable to chemical query
Adequate QSAR results
QSAR results relevant to regulatory purposes

The conditions for the validity of QSARs

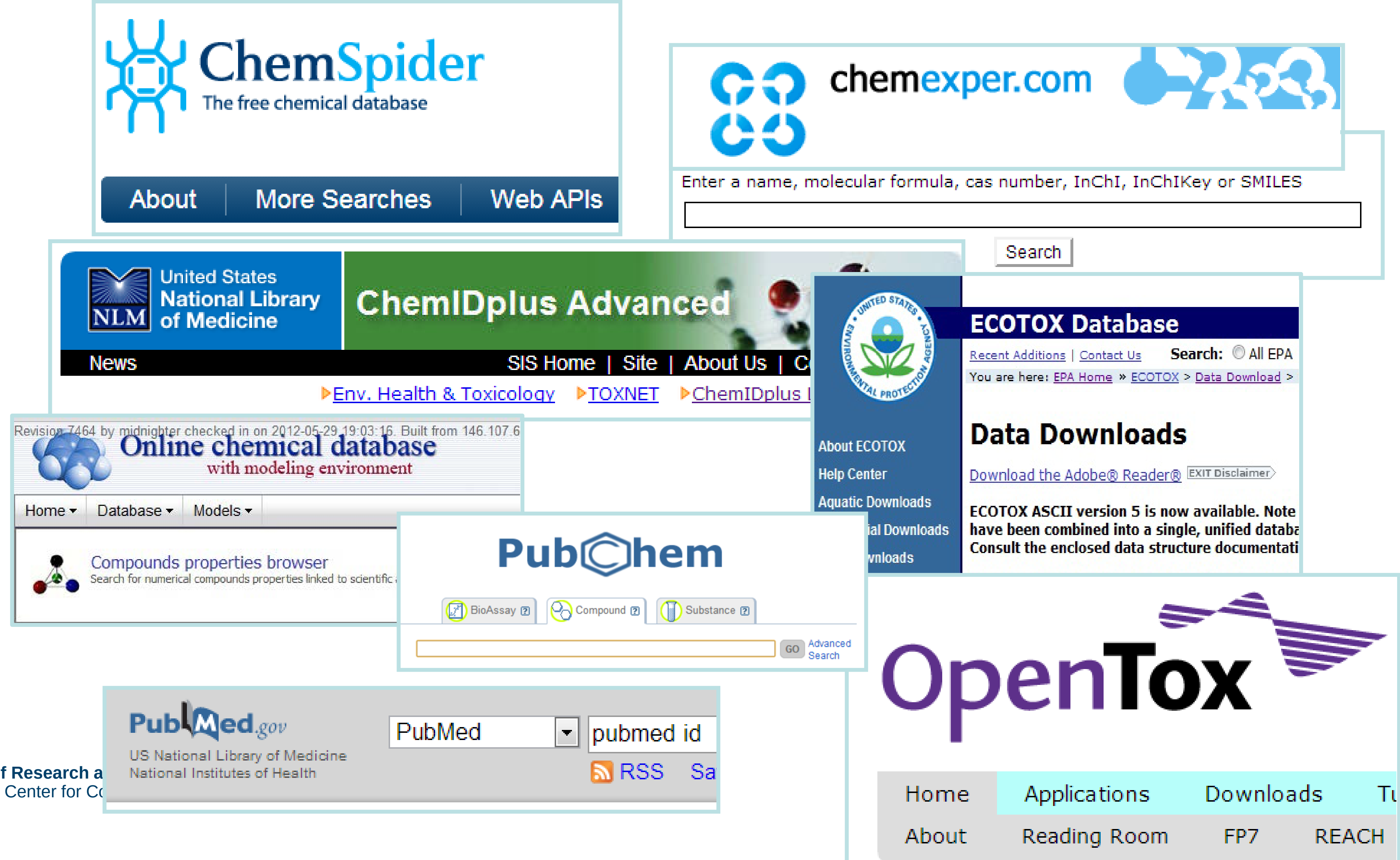
The 5 OECD principles:

Principle	Description
1) A defined endpoint	Any **physicochemical, biological or environmental** effect that can be measured and therefore modelled.
2) An unambiguous algorithm	**Ensure transparency** in the description of the model algorithm.
3) A defined domain of applicability	**Define limitations** in terms of the types of **chemical structures**, physicochemical properties and mechanisms of action for which the models can generate **reliable predictions**.
4) Appropriate measures of goodness-of-fit, robustness and predictivity	a) The internal **fitting** performance of a model b) the **predictivity** of a model, determined by using an appropriate **external test set**.
5) Mechanistic interpretation, if possible	Mechanistic **associations** between the **descriptors** used in a model and the **endpoint being predicted**.

Development of a QSAR model

- Curation of the data
  - » *Flagged and curated files available for sharing*
- Preparation of training and test sets
  - » *Inserted as a field in SDFiles and csv data files*
- Calculation of an initial set of descriptors
  - » *PaDEL 2D descriptors and fingerprints generated and shared*
- Selection of a mathematical method
  - » *Several approaches tested: KNN, PLS, SVM…*
- Variable selection technique
  - » *Genetic algorithm*
- Validation of the model’s predictive ability
  - » *5-fold cross validation and external test set*
- Define the Applicability Domain
  - » *Local (nearest neighbors) and global (leverage) approaches*

Public domain data sources



ChemSpider
The free chemical database
About | More Searches | Web APIs

chemexper.com
Enter a name, molecular formula, cas number, InChI, InChIKey or SMILES
Search

United States National Library of Medicine NLM
ChemIDplus Advanced
News | SIS Home | Site | About Us | C
Env. Health & Toxicology | TOXNET | ChemIDplus I

ECOTOX Database
Recent Additions | Contact Us | Search: ○ All EPA
You are here: EPA Home » ECOTOX > Data Download >
About ECOTOX
Help Center
Aquatic Downloads
ial Downloads
wnloads

Data Downloads

Download the Adobe® Reader® EXIT Disclaimer

ECOTOX ASCII version 5 is now available. Note
have been combined into a single, unified databa
Consult the enclosed data structure documentati

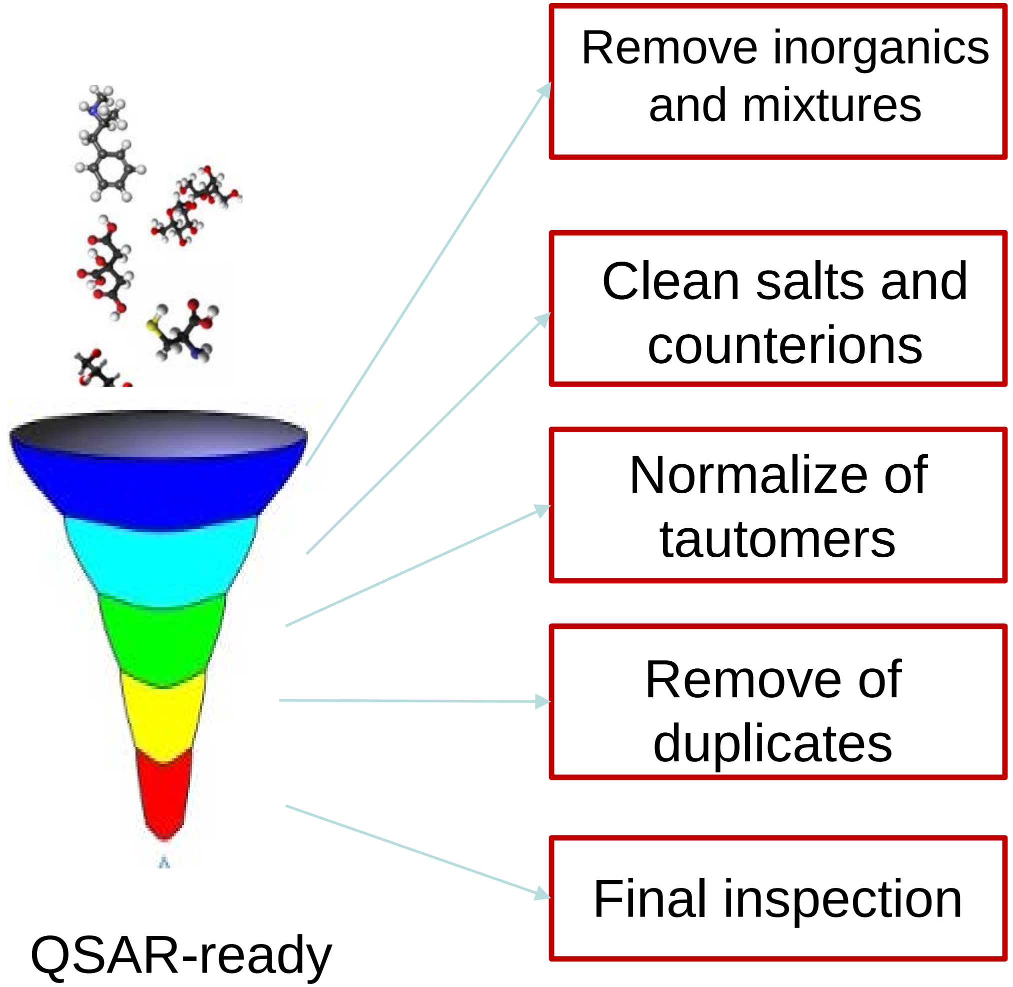
Revision 7464 by midnighter checked in on 2012-05-29 19:03:16. Built from 146.107.6
Online chemical database
with modeling environment
Home ▾ | Database ▾ | Models ▾
Compounds properties browser
Search for numerical compounds properties linked to scientific

PubChem
BioAssay | Compound | Substance
GO | Advanced Search

OpenTox
Home | Applications | Downloads | Tu
About | Reading Room | FP7 | REACH

PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed | pubmed id
RSS | Sa

Structure curation procedure



Remove inorganics and mixtures

Clean salts and counterions

Normalize of tautomers

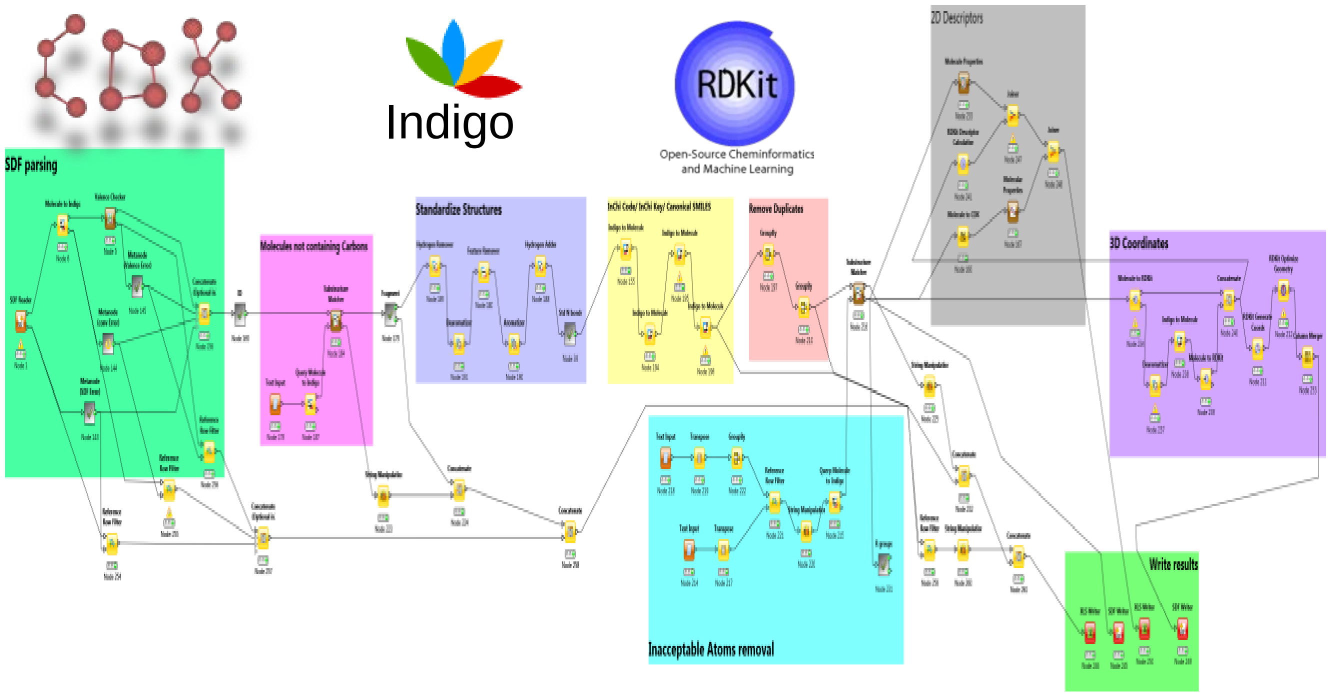
Remove of duplicates

Final inspection

QSAR-ready structures

Aim of the workflow:

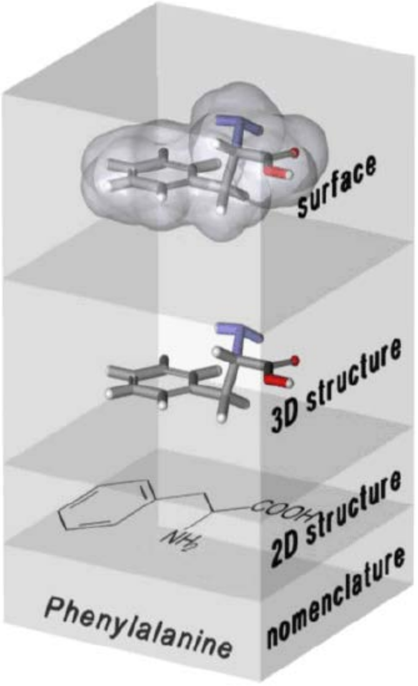
- Combine different procedures and ideas
- Minimize the differences between the structures used for prediction
- Produce a flexible free and open source workflow to be shared

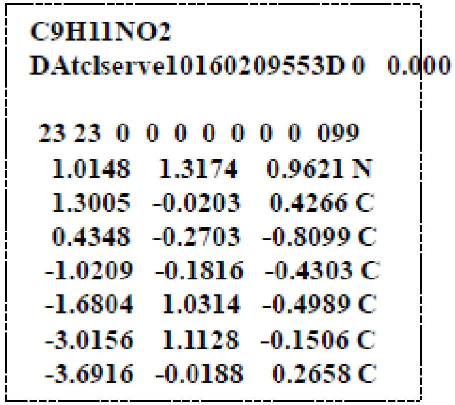

CDK
Indigo
RDKit
Open-Source Cheminformatics and Machine Learning
SDF parsing
Molecules not containing Carbons
Standardize Structures
InChI Code/ InChI Key/ Canonical SMILES
Remove Duplicates
2D Descriptors
3D Coordinates
Inacceptable Atoms removal
Write results

KNIME workflow

Mansouri et al. (2016) EHP 124:1023–1033 DOI:10.1289/ehp.1510267

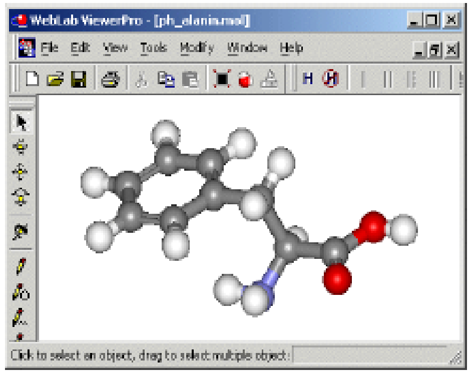
Molecular structures in the computer





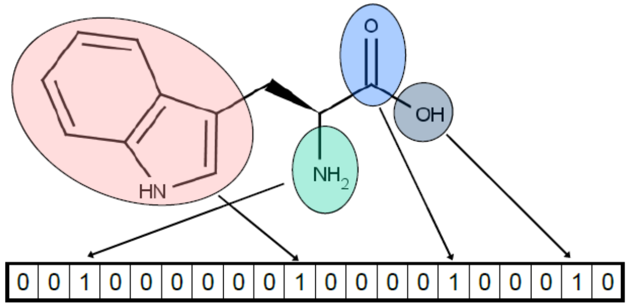
```
C9H11NO2
DAtclserve10160209553D 0   0.000

 23 23  0  0  0  0  0  0  0  099
  1.0148   1.3174   0.9621 N
  1.3005  -0.0203   0.4266 C
  0.4348  -0.2703  -0.8099 C
 -1.0209  -0.1816  -0.4303 C
 -1.6804   1.0314  -0.4989 C
 -3.0156   1.1128  -0.1506 C
 -3.6916  -0.0188   0.2658 C
```



Click to select an object, drag to select multiple object:

Bitstrings in databases



Fragmental keys & fingerprints

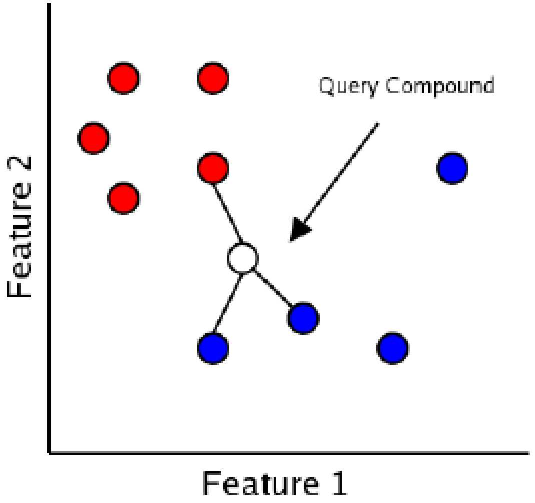
- substructural search

- read-across

- similarity search

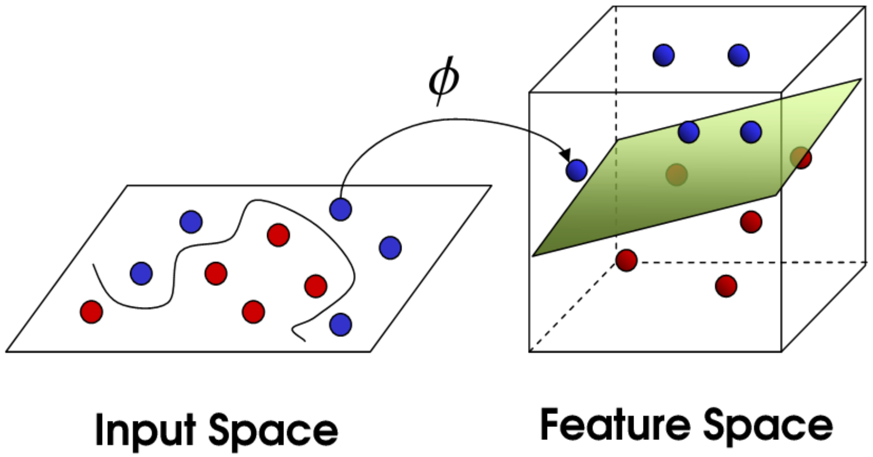
Classification methods

- ***k*NN: *k* Nearest Neighbors**



classification according to the majority class of the *k* neighbors

- **SVM: Support Vector Machines**



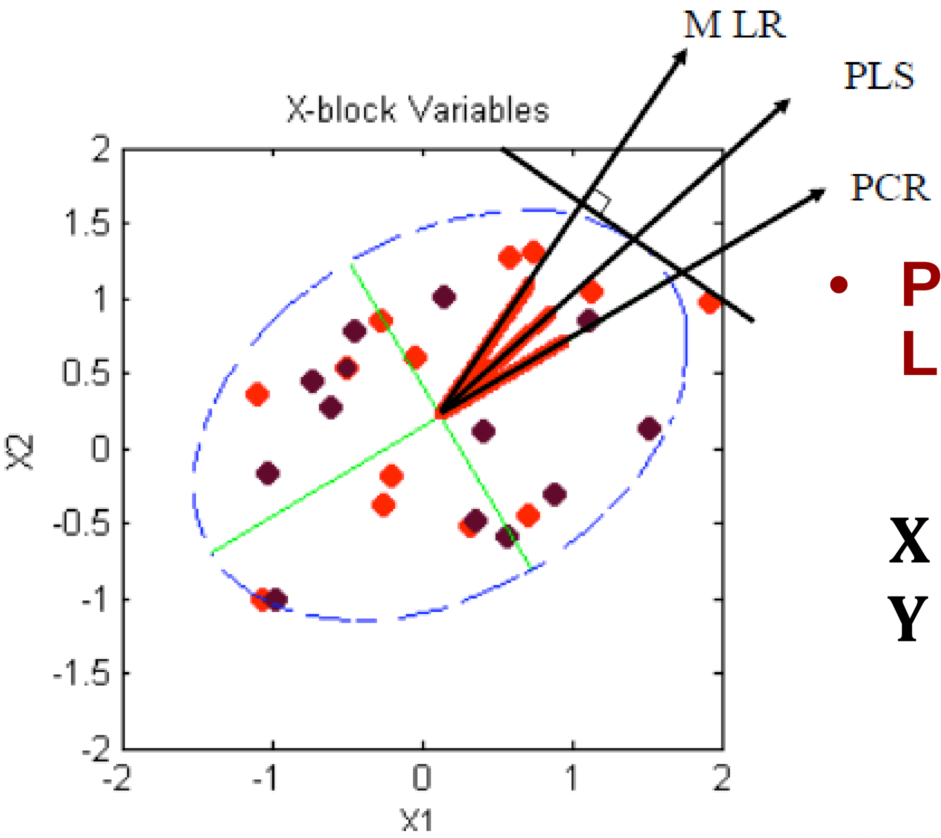
Kernel function maximizing the margin between the classes

Other methods: Self organized maps (SOM), Kohonen maps, PLSDA, LDA

Regression methods

- **MLR: Multiple Linear Regression**

$$\hat{y} = \mathbf{bX}$$
$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$



- **PLS: Partial Least Squares**

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E}$$
$$\mathbf{Y} = \mathbf{UQ}' + \mathbf{F}$$

PLS is the vector on the PCR ellipse upon which MLR has the longest projection

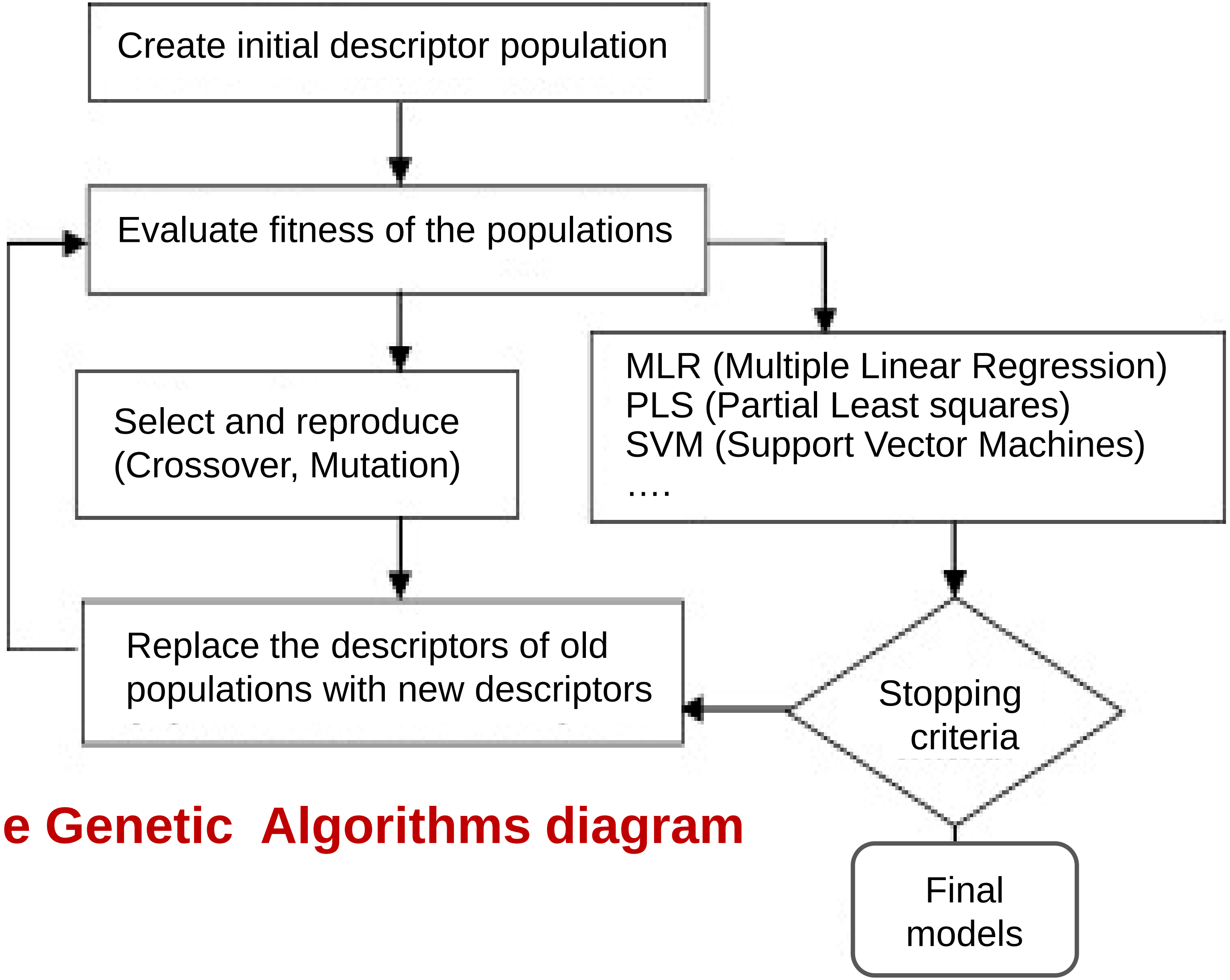
Other methods: Artificial Neural Networks (ANN), Random Forest, LASSO, PCR…

Variable selection procedure

- Many more descriptors than chemicals

- Many irrelevant descriptors

Only the most important descriptors are selected



Create initial descriptor population

Evaluate fitness of the populations

Select and reproduce (Crossover, Mutation)

MLR (Multiple Linear Regression)
PLS (Partial Least squares)
SVM (Support Vector Machines)
….

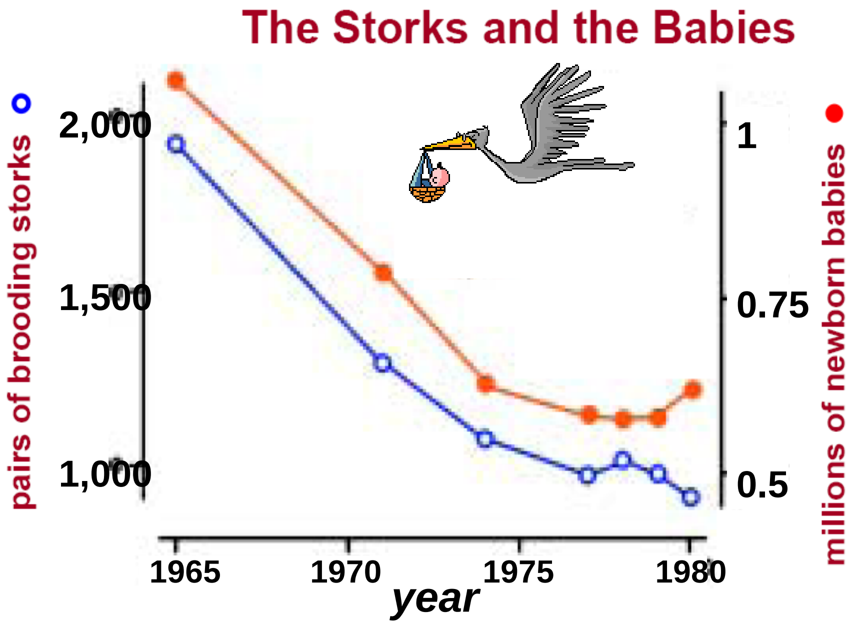
Replace the descriptors of old populations with new descriptors

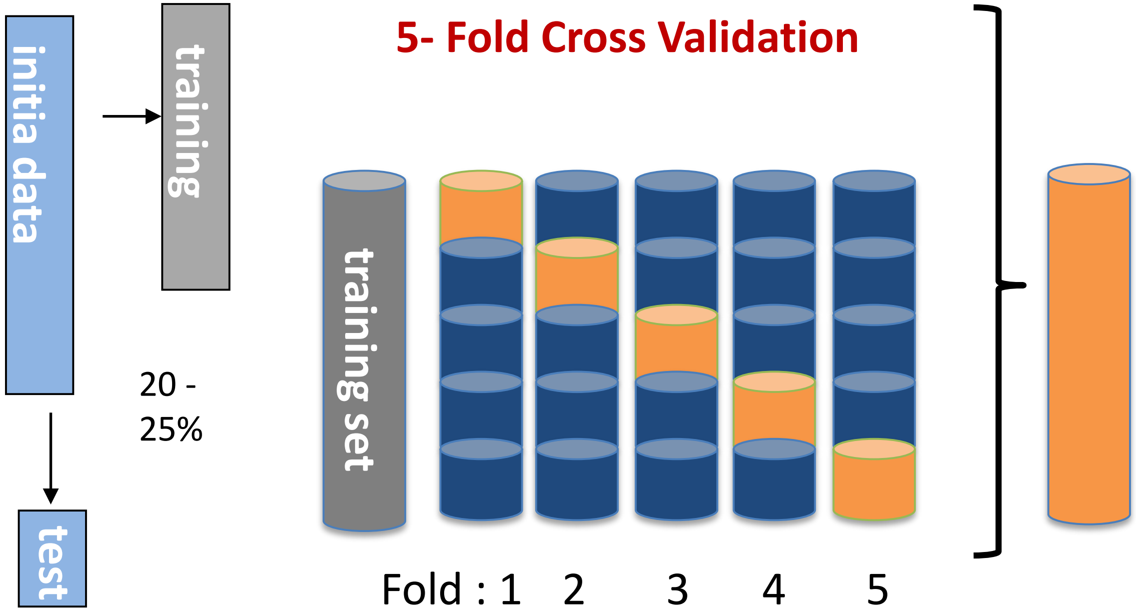
Stopping criteria

Final models

The Genetic Algorithms diagram

Cross-validation and test-set to avoid the “by chance” correlation problem

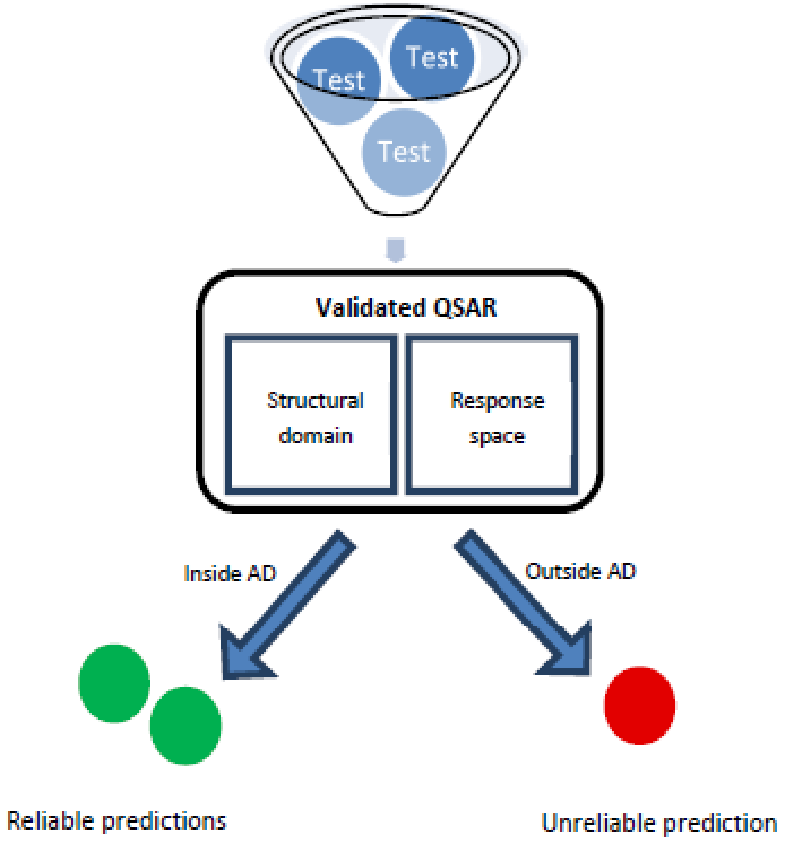


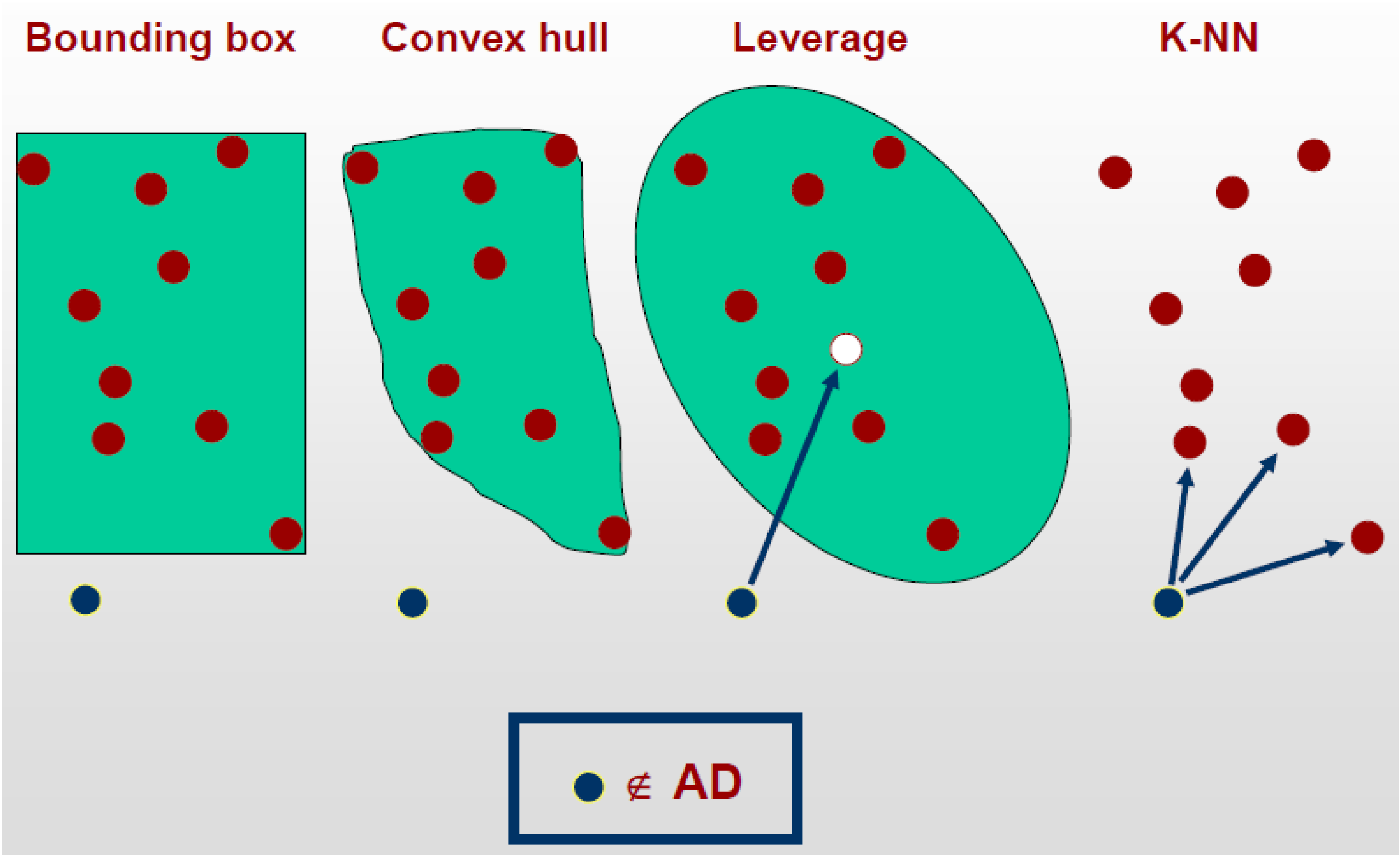


“There is a concern in West Germany over the falling **birth rate**. The accompanying graph might suggest a solution that **every child knows makes sense**”.

H. Sies, Nature 332, 495 (1988)

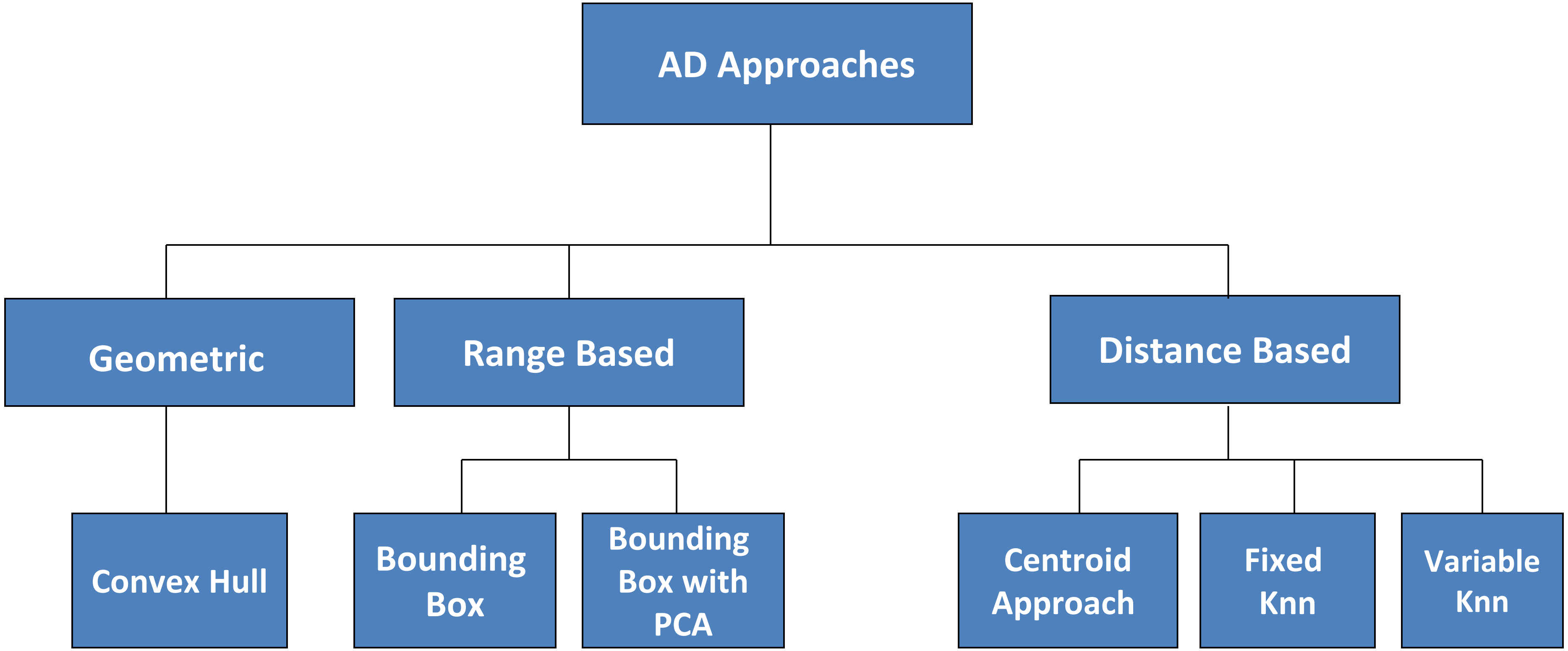
Defining the Applicability Domain (AD)





Sahigara, Mansouri et al. Molecules 17 (5), 4791-4810

An overview of Different AD Approaches



Sahigara, Mansouri et al. Molecules 17 (5), 4791-4810

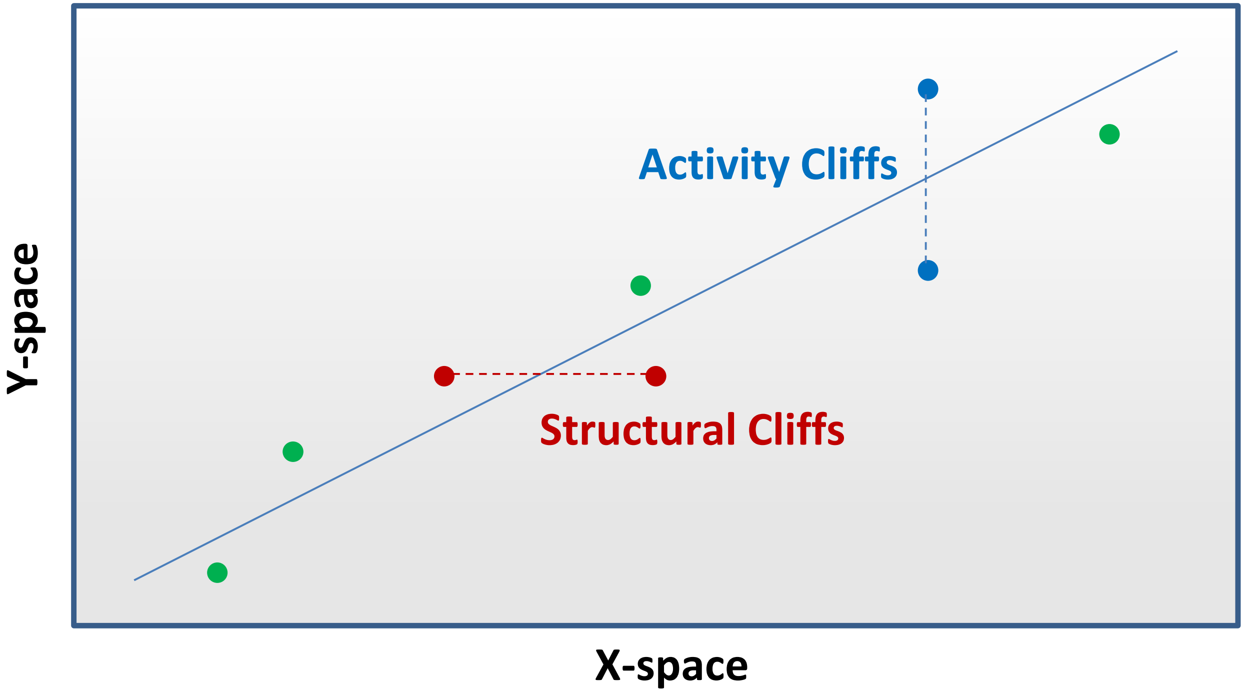
Structure-Activity landscape

Smooth landscape:
Congenericity principle fulfilled

Rugged landscape:
Activity cliffs & structural cliffs

Maggiora (2006): The difference between “*the gently rolling hills found on the Kansas prairie*” and “*the rugged landscapes of Utah’s Bryce Canyon*”

Activity cliffs/Structural cliffs



Activity Cliffs:

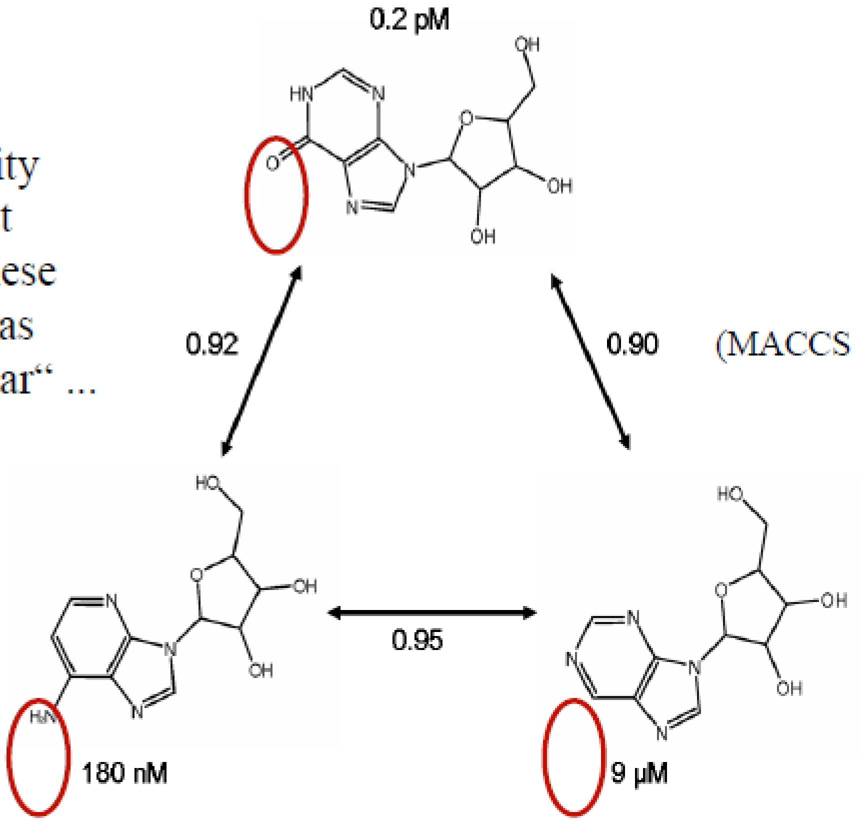
Two structurally similar compounds with diverse values of the activity

Structural Cliffs:

Two structurally diverse compounds with similar values of the activity

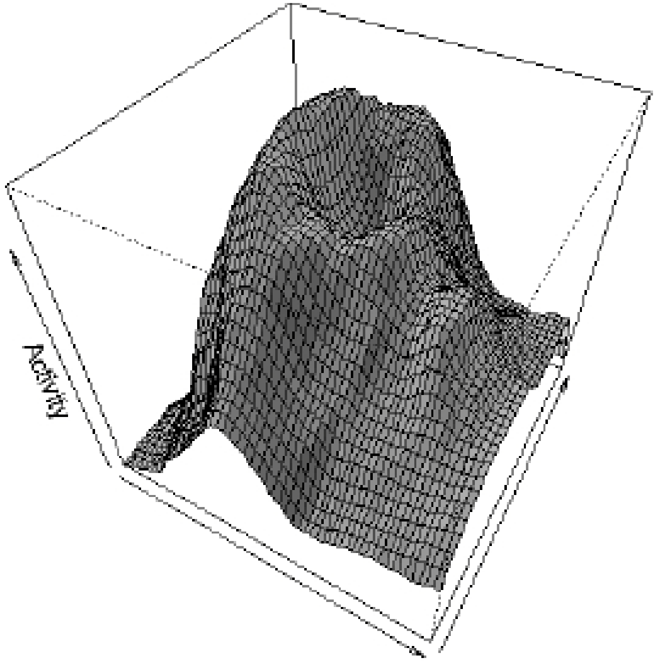
Discontinuous SARs

Any similarity method must recognize these compounds as being “similar“ …



(MACCS Tanimoto similarity)

Adenosine deaminase inhibitors



Structure-Activity Landscape index (SALI)

$$SALI_{st} = \frac{|A_s - A_t|}{1.01 - sim(s,t)}$$

A: the activity of a given molecule
Sim: the similarity coefficient

ER & AR modeling projects: Background and Goals

- U.S. Congress mandated that the EPA screen chemicals for their potential to be endocrine disruptors

- This led to the development of the Endocrine Disruptor Screening Program (EDSP)

- The initial focus was on environmental estrogens, but the program was expanded to include androgens and thyroid pathway disruptors

CERRAP : Collaborative Estrogen Receptor Activity Prediction Project

40 scientists, 17 research groups

- **EPA/NCCT:** U.S. Environmental Protection Agency / National Center for Computational Toxicology. **USA**
- **DTU/food:** Technical University of Denmark/ National Food Institute. **Denmark**
- **FDA/NCTR/DBB:** U.S. Food and Drug Administration. **USA**
- **FDA/NCTR/DSB:** U.S. Food and Drug Administration. **USA**
- **Helmholtz/ISB:** Helmholtz Zentrum Muenchen/Institute of Structural Biology. **Germany**
- **ILS&EPA/NCCT:** ILS Inc & EPA/NCCT. **USA**
- **IRCSS:** Istituto di Ricerche Farmacologiche “Mario Negri”. **Italy**
- **JRC_Ispra:** Joint Research Centre of the European Commission, Ispra. **Italy**
- **LockheedMartin&EPA:** Lockheed Martin IS&GS/ High Performance Computing. **USA**
- **NIH/NCATS:** National Institutes of Health/ National Center for Advancing Translational Sciences. **USA**
- **NIH/NCI:** National Institutes of Health/ National Cancer Institute. **USA**
- **RIFM:** Research Institute for Fragrance Materials, Inc. **USA**
- **UMEA/Chemistry:** University of UMEA/ Chemistry department. **Sweden**
- **UNC/MML:** University of North Carolina/ Laboratory for Molecular Modeling. **USA**
- **UniBA/Pharma:** University of Bari/ Department of Pharmacy. **Italy**
- **UNIMIB/Michem:** University of Milano-Bicocca/ Milano Chemometrics and QSAR Research Group. **Italy**
- **UNISTRA/Infochim:** University of Strasbourg/ ChemoInformatique. **France**

Mansouri et al. (2016) EHP 124:1023–1033 DOI:10.1289/ehp.1510267

United States Environmental Protection Agency

CERAPP data and results

Datasets of the project

- Training set: 1,677 chemicals **(EPA ToxCast data)**
- Prediction set: 32,464 chemicals **(The Human Exposure Universe)**
- Evaluation set: 7,000 chemicals **(Literature: Tox21, FDA, METI…)**

40 Models received:

- Classification / Qualitative:
  - Binding: **22 models**
  - Agonists: **11 models**
  - Antagonists: **9 models**

Regression / Quantitative:

Binding: **3 models**

Agonists: **3 models**

Antagonists: **2 models**

Consensus modeling:

Weighted vote based on rankings of the predictions accuracy scores

Mansouri et al. (2016) EHP 124:1023–1033 DOI:10.1289/ehp.1510267

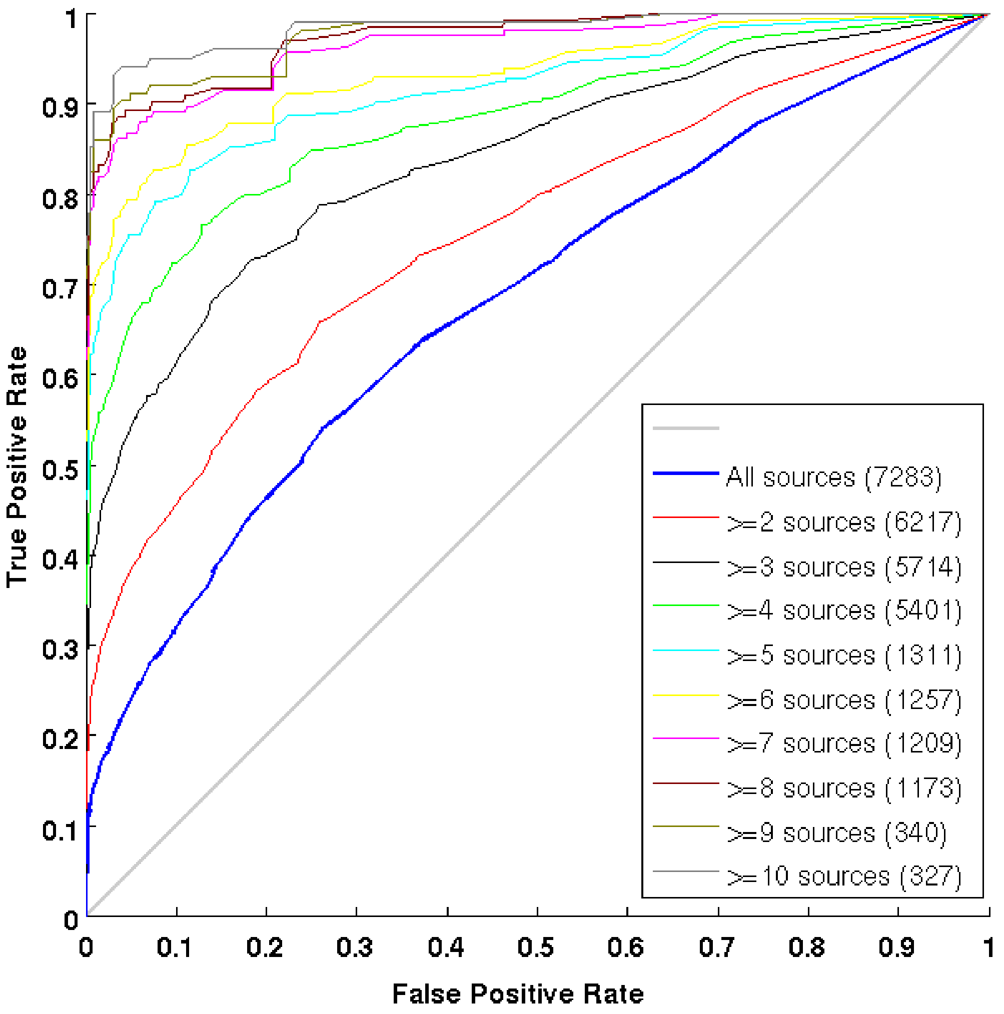
Consensus Qualitative Accuracy

Prediction Accuracy Strongly Depends on Data Quality

Total binders: **3961**
Agonists: **2494**
Antagonists: **2793**

	ToxCast data (training set)		Literature data (test set)	
Observed\Predicted	**Actives**	**Inactives**	**Actives**	**Inactives**
Actives	83	6	597	1385
Inactives	40	1400	463	4838

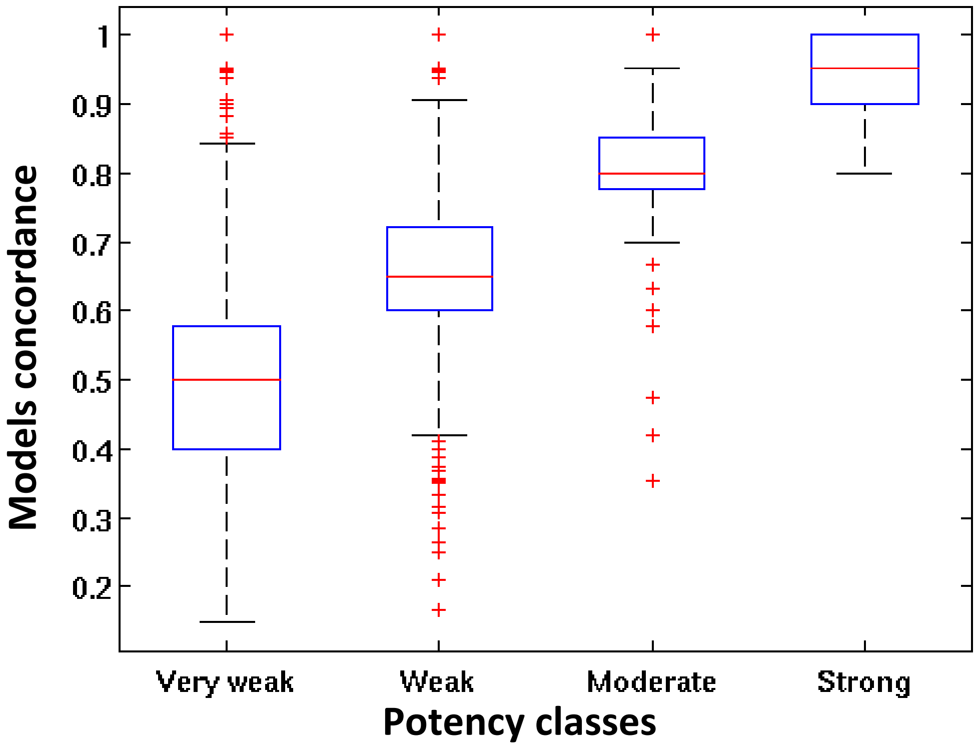
	ToxCast data	Literature data (All: 7283)	Literature data (>6 sources: 1209)
Sensitivity	0.93	0.30	0.87
Specificity	0.97	0.91	0.94
Balanced accuracy	0.95	0.61	0.91



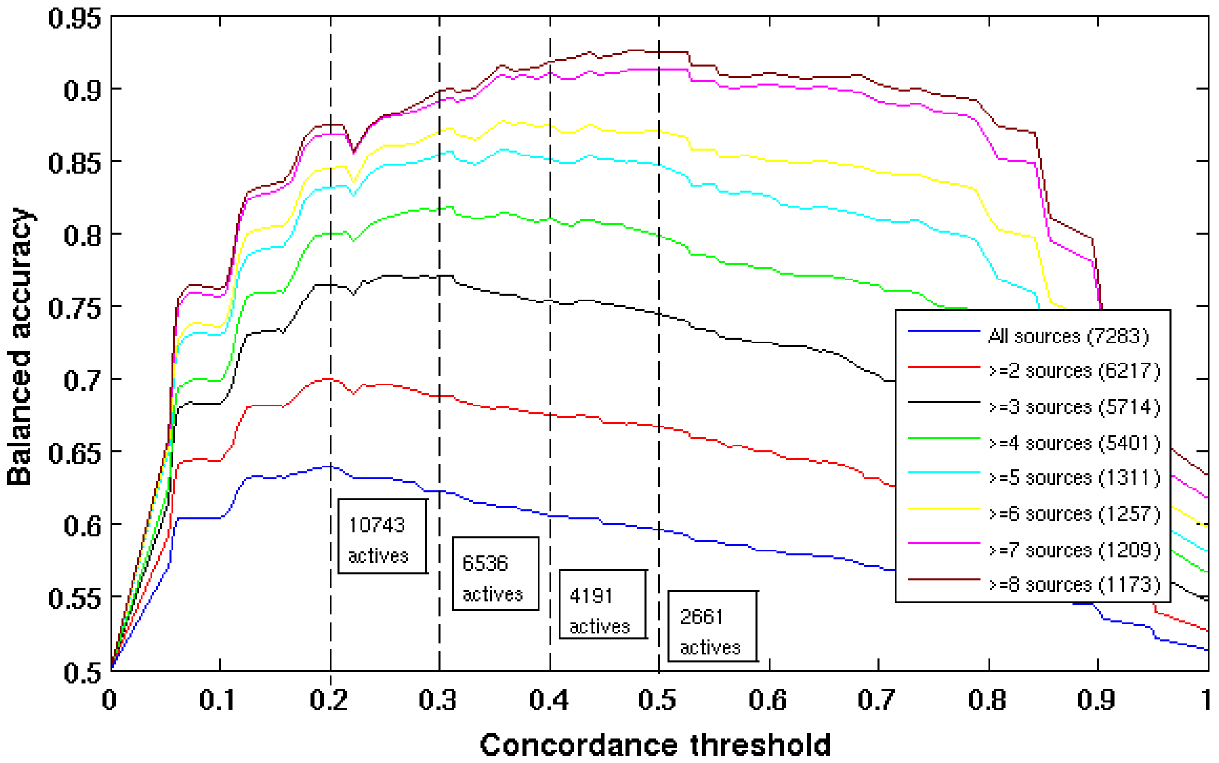
ROC curve of the external validation set (literature)

Mansouri et al. (2016) EHP 124:1023–1033 DOI:10.1289/ehp.1510267

Consensus Quantitative Accuracy



Box plot of the active classes of the consensus model.



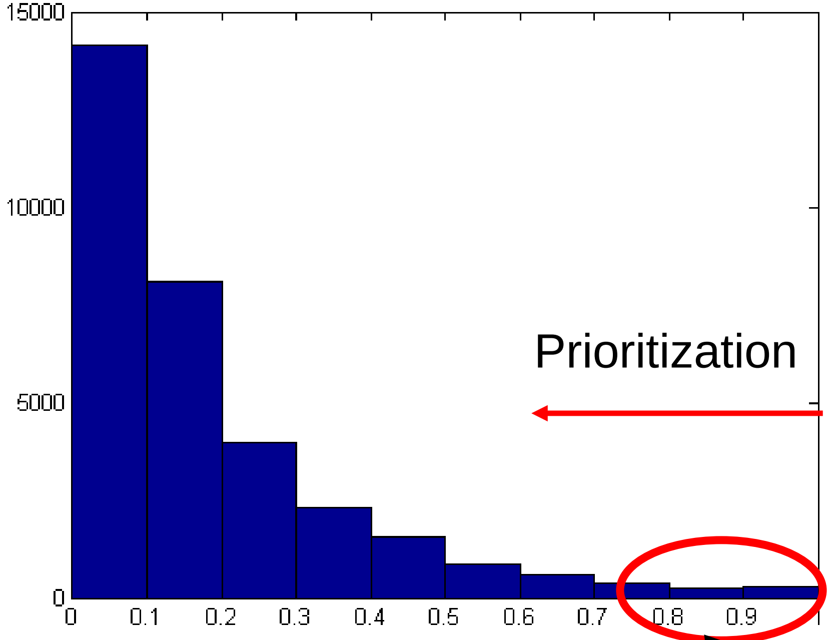
Variation of the balanced accuracy with positive concordance thresholds

- positive concordance < 0.6 => Potency class= **Very weak**
- 0.6=<positive concordance<0.75 => Potency class= **Weak**
- 0.75=<positive concordance<0.9 => Potency class= **Moderate**
- positive concordance>=0.9 => Potency class= **Strong**

Mansouri et al. (2016) EHP 124:1023–1033 DOI:10.1289/ehp.1510267

Concordance of the qualitative models

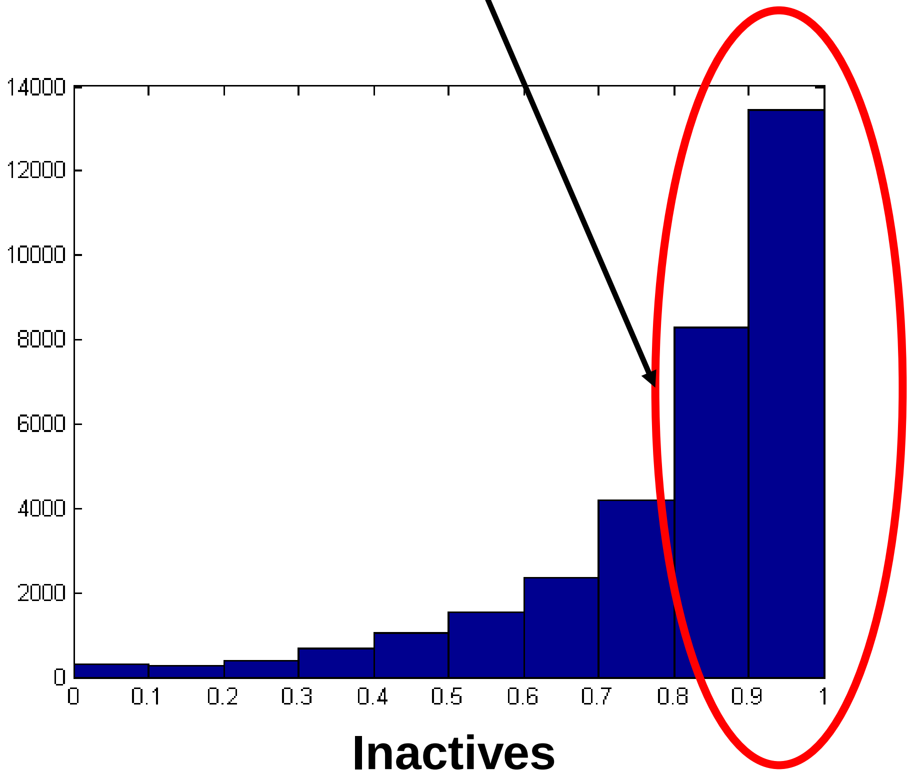
Actives



Prioritization

Only 757 chemicals have >75% positive concordance

Most models predict most chemicals as inactive



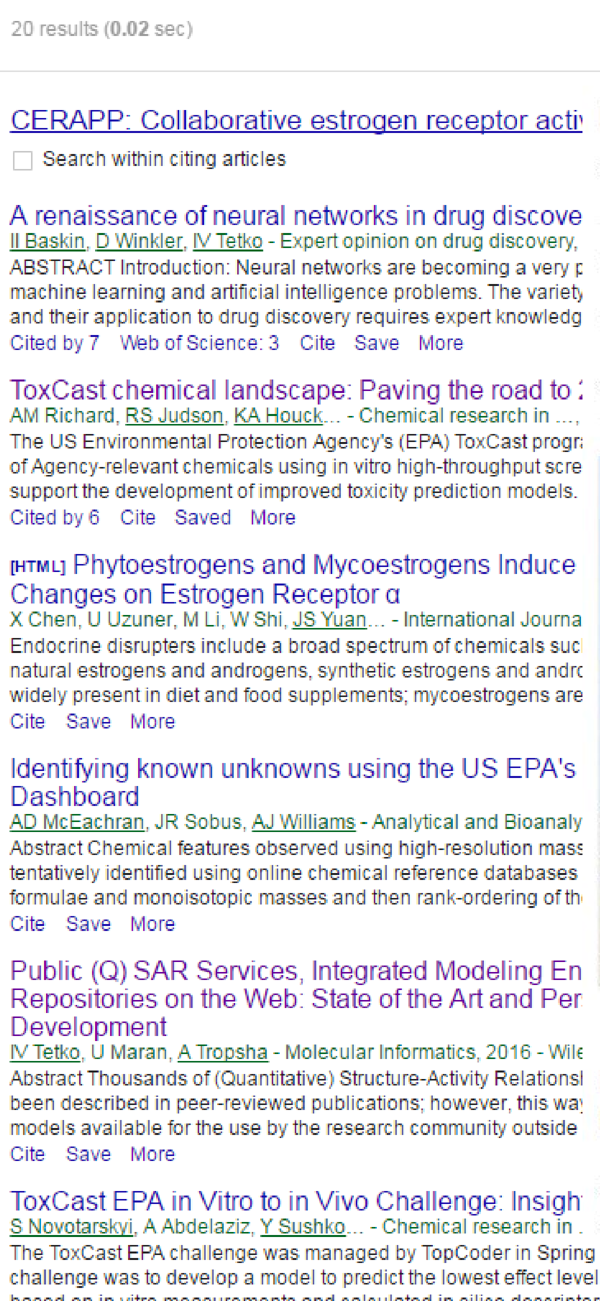
Inactives

⇨ Only a small fraction of chemicals require further testing!

Mansouri et al. (2016) EHP 124:1023–1033 DOI:10.1289/ehp.1510267

Mansouri et al. (2016) EHP 124:1023–1033 DOI:10.1289/ehp.1510267

ehp ENVIRONMENTAL HEALTH PERSPECTIVES



20 results (**0.02** sec)

CERAPP: Collaborative estrogen receptor acti

□ Search within citing articles

A renaissance of neural networks in drug discove
Il Baskin, D Winkler, IV Tetko - Expert opinion on drug discovery,
ABSTRACT Introduction: Neural networks are becoming a very p
machine learning and artificial intelligence problems. The variety
and their application to drug discovery requires expert knowledg
Cited by 7 Web of Science: 3 Cite Save More

ToxCast chemical landscape: Paving the road to 2
AM Richard, RS Judson, KA Houck… - Chemical research in …,
The US Environmental Protection Agency's (EPA) ToxCast progr
of Agency-relevant chemicals using in vitro high-throughput scre
support the development of improved toxicity prediction models.
Cited by 6 Cite Saved More

[HTML] Phytoestrogens and Mycoestrogens Induce Changes on Estrogen Receptor α
X Chen, U Uzuner, M Li, W Shi, JS Yuan… - International Journa
Endocrine disruptors include a broad spectrum of chemicals suc
natural estrogens and androgens, synthetic estrogens and andr
widely present in diet and food supplements; mycoestrogens are
Cite Save More

Identifying known unknowns using the US EPA's Dashboard
AD McEachran, JR Sobus, AJ Williams - Analytical and Bioanaly
Abstract Chemical features observed using high-resolution mass
tentatively identified using online chemical reference databases
formulae and monoisotopic masses and then rank-ordering of th
Cite Save More

Public (Q) SAR Services, Integrated Modeling En Repositories on the Web: State of the Art and Per Development
IV Tetko, U Maran, A Tropsha - Molecular Informatics, 2016 - Wile
Abstract Thousands of (Quantitative) Structure-Activity Relationsh
been described in peer-reviewed publications; however, this wa
models available for the use by the research community outside
Cite Save More

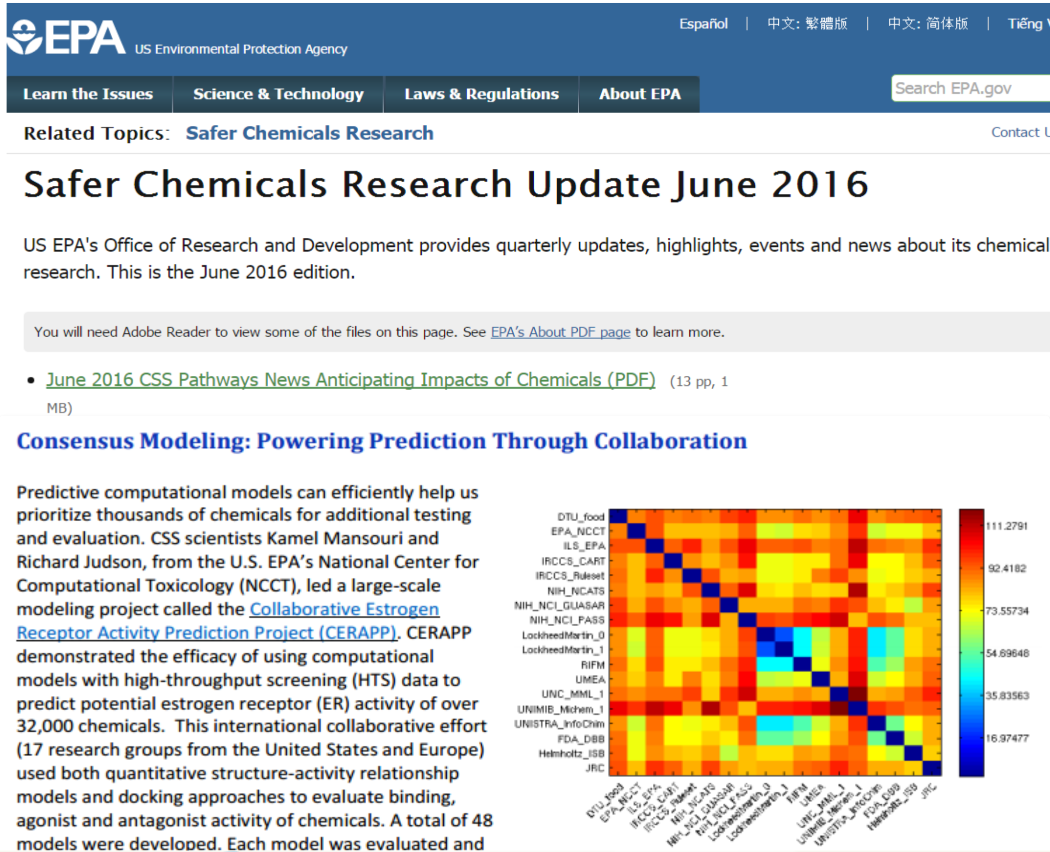
ToxCast EPA in Vitro to in Vivo Challenge: Insigh
S Novotarskyi, A Abdelaziz, Y Sushko… - Chemical research in
The ToxCast EPA challenge was managed by TopCoder in Spring 2014. The goal of the
challenge was to develop a model to predict the lowest effect level (LEL) concentration
based on in vitro measurements and calculated in silico descriptors. This article summarizes

US Government Information

One stop source for US Government Information

EDSP Prioritization: Collaborative Estrogen Recepto Prediction Project (CERAPP) (SOT)

Humans are potentially exposed to tens of thousands of man-made chemicals in
environment. It is well known that some environmental chemicals mimic natural



Related Topics: Safer Chemicals Research

Safer Chemicals Research Update June 2016

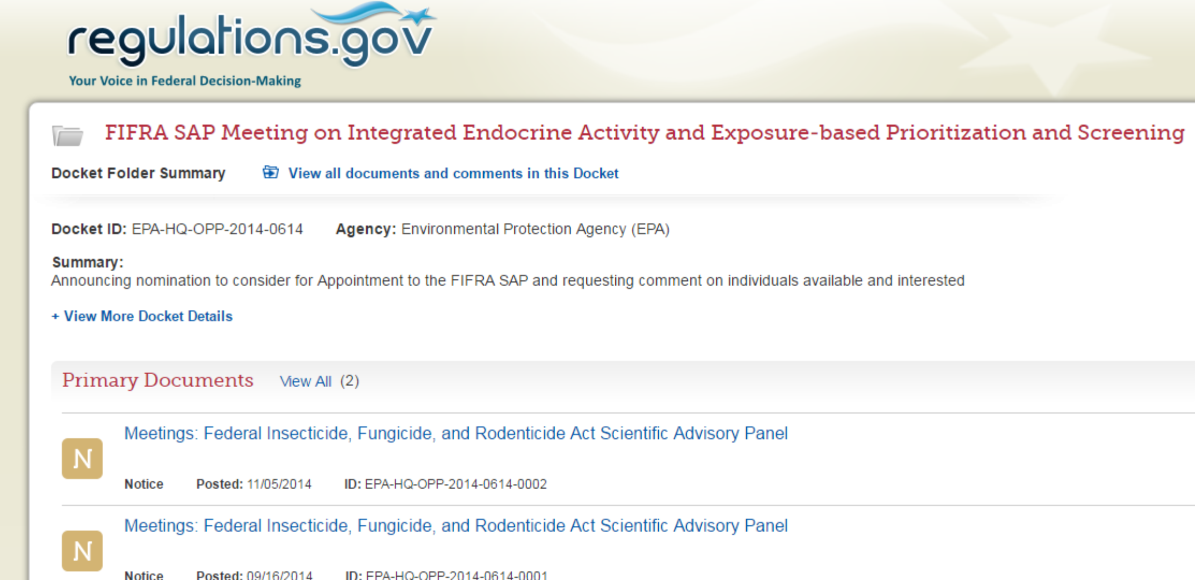
US EPA's Office of Research and Development provides quarterly updates, highlights, events and news about its chemical research. This is the June 2016 edition.

You will need Adobe Reader to view some of the files on this page. See EPA's About PDF page to learn more.

- June 2016 CSS Pathways News Anticipating Impacts of Chemicals (PDF) (13 pp, 1 MB)

Consensus Modeling: Powering Prediction Through Collaboration

Predictive computational models can efficiently help us prioritize thousands of chemicals for additional testing and evaluation. CSS scientists Kamel Mansouri and Richard Judson, from the U.S. EPA's National Center for Computational Toxicology (NCCT), led a large-scale modeling project called the Collaborative Estrogen Receptor Activity Prediction Project (CERAPP). CERAPP demonstrated the efficacy of using computational models with high-throughput screening (HTS) data to predict potential estrogen receptor (ER) activity of over 32,000 chemicals. This international collaborative effort (17 research groups from the United States and Europe) used both quantitative structure-activity relationship models and docking approaches to evaluate binding, agonist and antagonist activity of chemicals. A total of 48 models were developed. Each model was evaluated and



FIFRA SAP Meeting on Integrated Endocrine Activity and Exposure-based Prioritization and Screening

Docket Folder Summary   View all documents and comments in this Docket

Docket ID: EPA-HQ-OPP-2014-0614   **Agency:** Environmental Protection Agency (EPA)

Summary:
Announcing nomination to consider for Appointment to the FIFRA SAP and requesting comment on individuals available and interested

+ View More Docket Details

Primary Documents View All (2)

N Meetings: Federal Insecticide, Fungicide, and Rodenticide Act Scientific Advisory Panel
Notice **Posted:** 11/05/2014 **ID:** EPA-HQ-OPP-2014-0614-0002

N Meetings: Federal Insecticide, Fungicide, and Rodenticide Act Scientific Advisory Panel
Notice **Posted:** 09/16/2014 **ID:** EPA-HQ-OPP-2014-0614-0001

From CERAPP to CoMPARA : Collaborative **M**odeling **P**roject for **A**ndrogen **R**eceptor **A**ctivity

- Follow the CERAPP framework
- Use larger size prioritization set
- Use data from the combined EPA ToxCast AR assays
- Collect and curate data from the literature for validation
- Use agonists, antagonists, and binding data
- Build continuous and classification models
- Similar approach for consensus modeling and validation

CoMPARA participants: 34 international groups

From CERAPP

- **EPA/NCCT. USA**
- **DTU/food. Denmark**
- **FDA/NCTR/DBB. USA**
- **Helmholtz. Germany**
- **ILS&EPA/NCCT. USA**
- **IRCSS. Italy**
- **LockheedMartin&EPA. USA**
- **NIH/NCATS. USA**
- **NIH/NCI. USA**
- **UMEA/Chemistry. Sweden**
- **UNC/MML. USA**
- **UniBA/Pharma. Italy**
- **UNIMIB/Michem. Italy**
- **UNISTRA/Infochim. France**
- **VCCLab. Germany**

New research groups

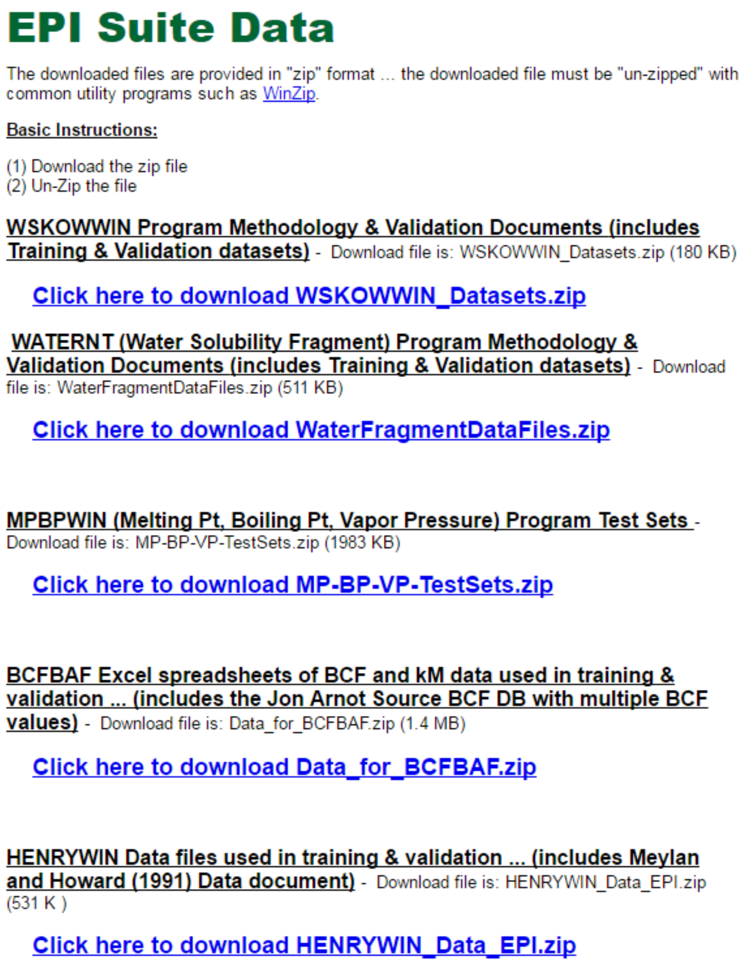
- **NCSU.** Department of Chemistry, Bioinformatics Research Center. **USA**
- **EPA/NRMRL.** National Risk Management Research Laboratory. **USA**
- **INSUBRIA.** University of Insubria**.** Environmental Chemistry. **Italy**
- **Tartu. University of Tartu.** Institute of Chemistry. **Estonia**
- **NIH/NTP/NICEATM. USA**
- **Chemistry Institute.** Lab of Chemometrics**. Slovenia**
- **SWETOX.** Swedish toxicology research center**. Sweden**
- **Lanzhou University . China**
- **BDS.** Biodetection Systems**. Netherlands**
- **MTI.** Molecules Theurapetiques in silico**. France**
- **IBMC.** Institute of Biomedical Chemistry. **Russia**
- **UNIMORE.** University of Modena Reggio-Emilia**. Italy**
- **UFG.** Federal University of Golas**. Brazil**
- **MSU.** Moscow State University**. Russia**
- **ZJU.** Zhejiang University. **China**
- **JKU.** Johannes Kepler University. **Austria**
- **CTIS.** Centre de Traitement de l'Information Scientifique. **France**
- **IdeaConsult. Bulgaria**
- **ECUST**. East China University of Science and Technology. **China**

Developing “OPERA Models”

- Interest in physicochemical properties to include in exposure modeling, augmented with ToxCast HTS *in vitro* data etc.
- Our approach to modeling:
  - Obtain high quality training sets
  - Apply appropriate modeling approaches
  - Validate performance of models
  - Define the applicability domain and limitations of the models
  - Use models to predict properties across our full datasets

PHYSPROP Data: Available from:

http://esc.syrres.com/interkow/EpiSuiteData.htm



EPI Suite Data

The downloaded files are provided in "zip" format ... the downloaded file must be "un-zipped" with common utility programs such as WinZip.

Basic Instructions:

(1) Download the zip file
(2) Un-Zip the file

WSKOWWIN Program Methodology & Validation Documents (includes Training & Validation datasets) - Download file is: WSKOWWIN_Datasets.zip (180 KB)

Click here to download WSKOWWIN_Datasets.zip

WATERNT (Water Solubility Fragment) Program Methodology & Validation Documents (includes Training & Validation datasets) - Download file is: WaterFragmentDataFiles.zip (511 KB)

Click here to download WaterFragmentDataFiles.zip

MPBPWIN (Melting Pt, Boiling Pt, Vapor Pressure) Program Test Sets - Download file is: MP-BP-VP-TestSets.zip (1983 KB)

Click here to download MP-BP-VP-TestSets.zip

BCFBAF Excel spreadsheets of BCF and kM data used in training & validation ... (includes the Jon Arnot Source BCF DB with multiple BCF values) - Download file is: Data_for_BCFBAF.zip (1.4 MB)

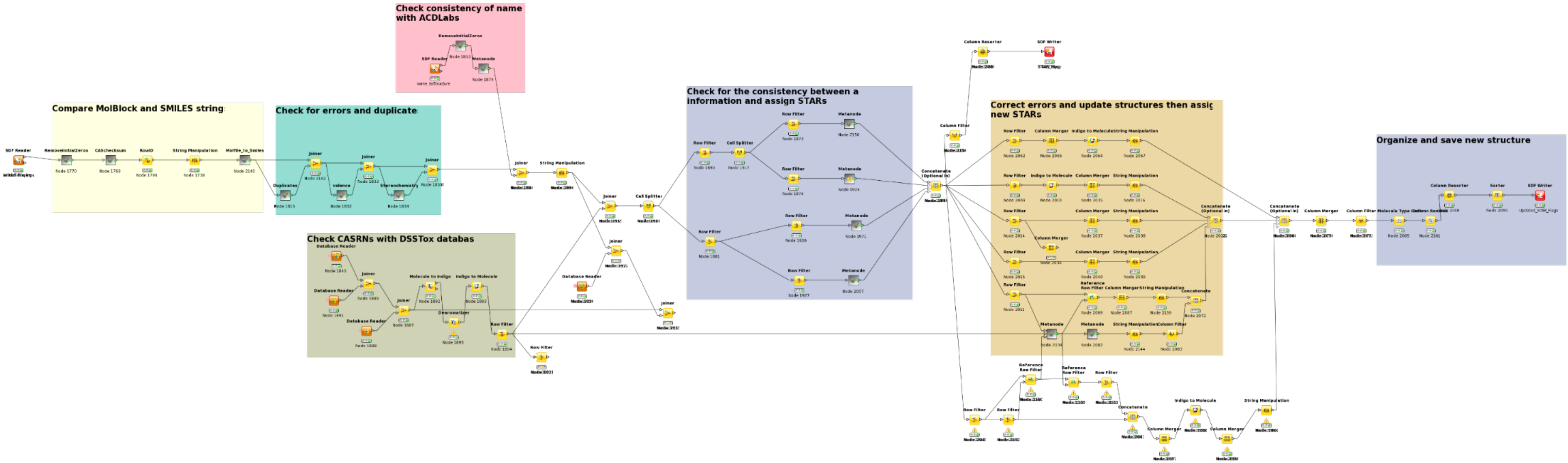
Click here to download Data_for_BCFBAF.zip

HENRYWIN Data files used in training & validation ... (includes Meylan and Howard (1991) Data document) - Download file is: HENRYWIN_Data_EPI.zip (531 K)

Click here to download HENRYWIN_Data_EPI.zip

- Water solubility
- Melting Point
- Boiling Point
- LogP (KOWWIN: Octanol-water partition coefficient)
- Atmospheric Hydroxylation Rate
- LogBCF (Bioconcentration Factor)
- Biodegradation Half-life
- Ready biodegradability
- Henry's Law Constant
- Fish Biotransformation Half-life
- LogKOA (Octanol/Air Partition Coefficient)
- LogKOC (Soil Adsorption Coefficient)
- Vapor Pressure

KNIME Workflow to Evaluate the Dataset



Mansouri et al. SAR QSAR Environ. Res. 2016, 27 (11), 939−965.

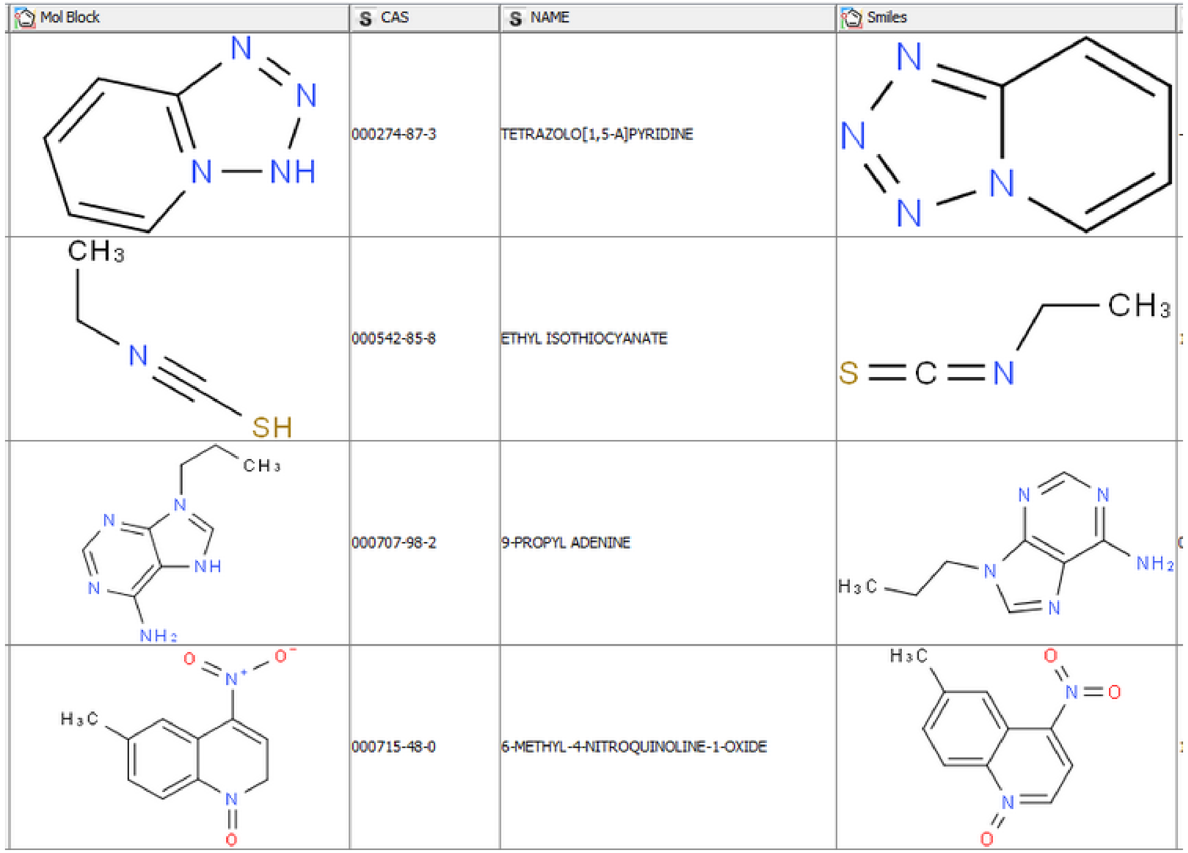
LogP dataset: 15,809 chemicals (structures)

- CAS Checksum: 12163 valid, 3646 invalid **(>23%)**
- Invalid names: 555
- Invalid SMILES 133
- Valence errors: 322 Molfile, 3782 SMILES **(>24%)**
- Duplicates check:
  - 31 DUPLICATE MOLFILES
  - 626 DUPLICATE SMILES
  - 531 DUPLICATE NAMES
- SMILES vs. Molfiles (structure check)
  - 1279 differ in stereochemistry **(~8%)**
  - 362 “Covalent Halogens”
  - 191 differ as tautomers
  - 436 are different compounds **(~3%)**

Mansouri et al. SAR QSAR Environ. Res. 2016, 27 (11), 939−965.

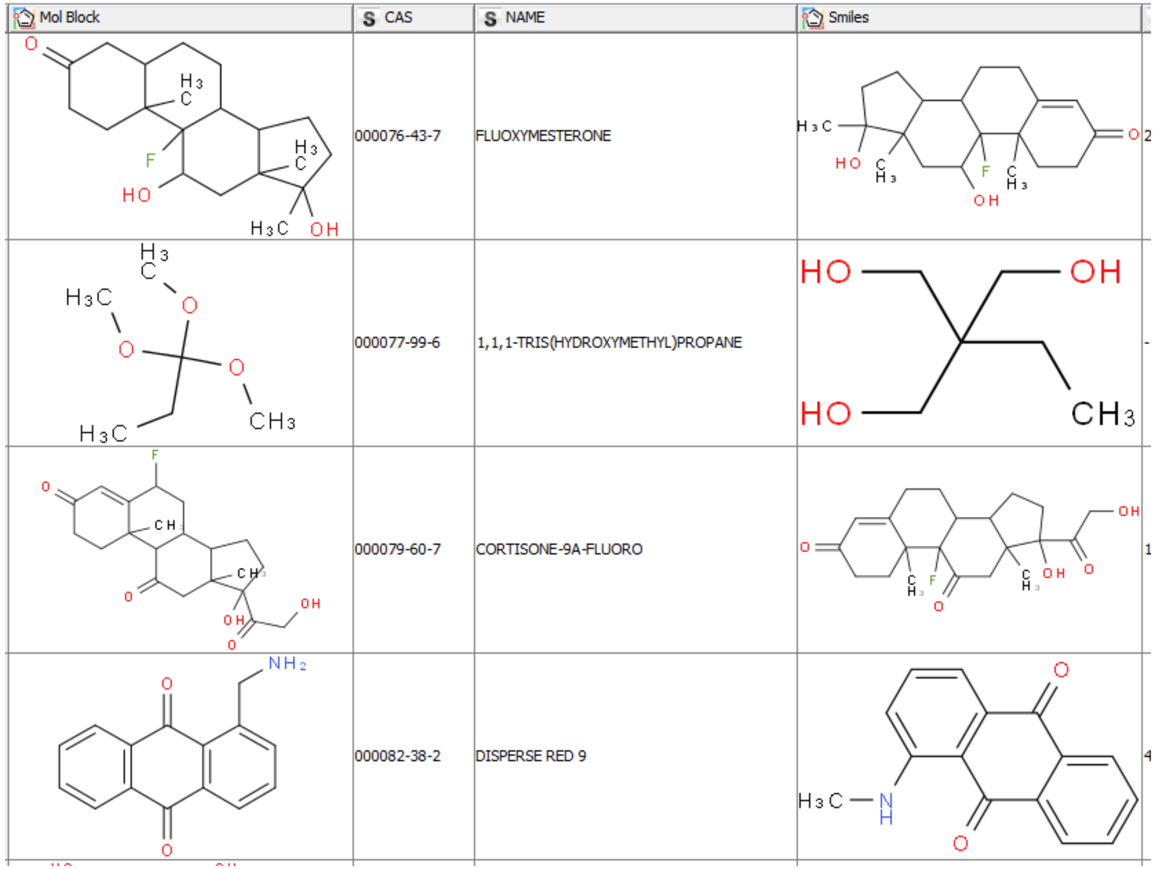
Examples of Errors

Valence Errors



Mol Block	CAS	NAME	Smiles
	000274-87-3	TETRAZOLO[1,5-A]PYRIDINE	
	000542-85-8	ETHYL ISOTHIOCYANATE	
	000707-98-2	9-PROPYL ADENINE	
	000715-48-0	6-METHYL-4-NITROQUINOLINE-1-OXIDE	

Different Compounds

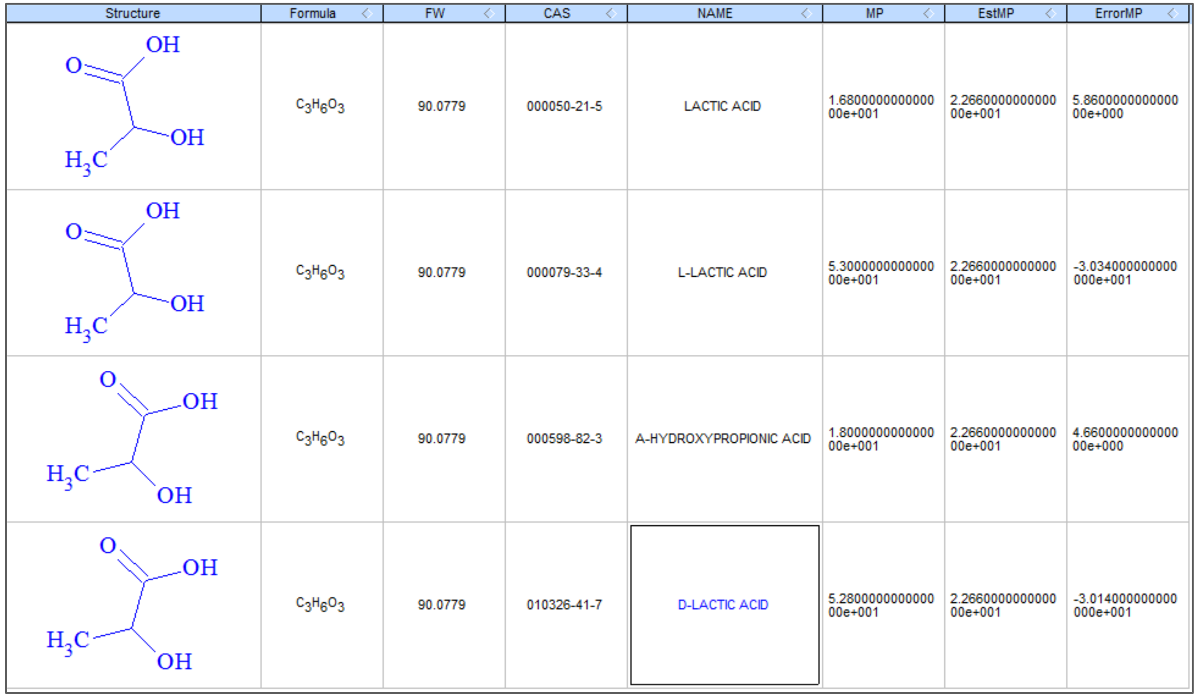


Mol Block	CAS	NAME	Smiles
	000076-43-7	FLUOXYMESTERONE	
	000077-99-6	1,1,1-TRIS(HYDROXYMETHYL)PROPANE	
	000079-60-7	CORTISONE-9A-FLUORO	
	000082-38-2	DISPERSE RED 9	

Mansouri et al. SAR QSAR Environ. Res. 2016, 27 (11), 939−965.

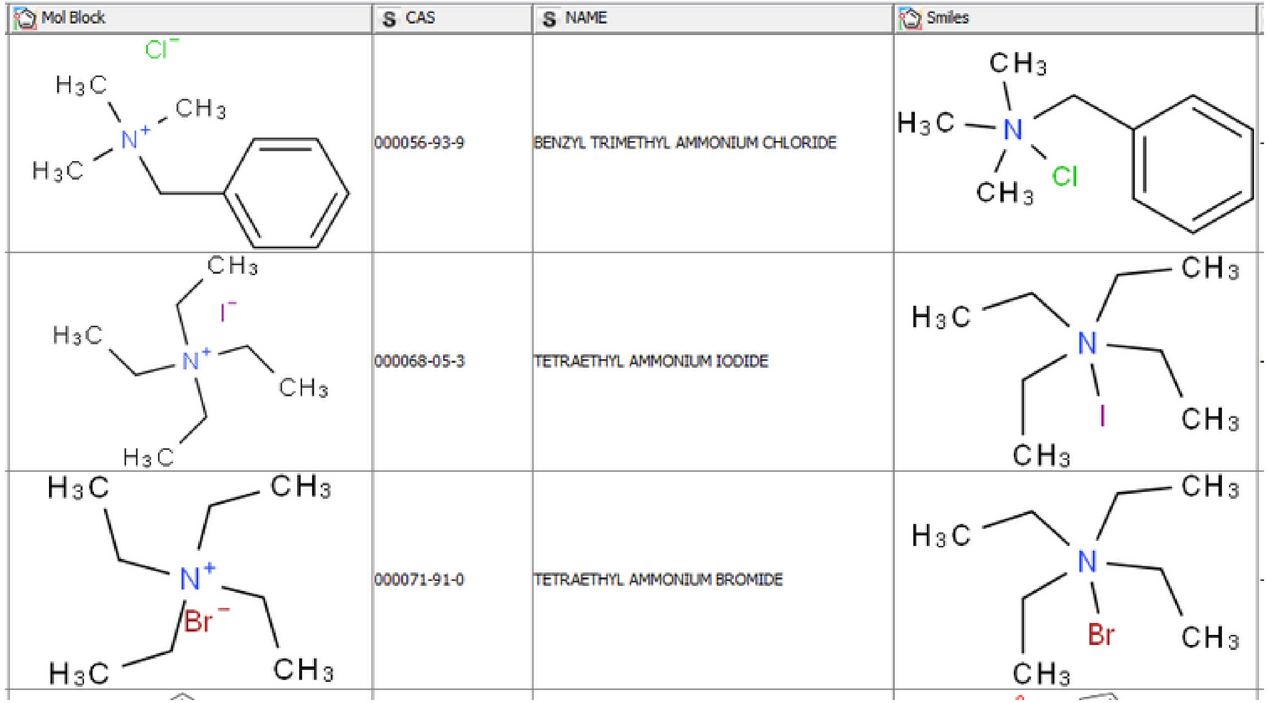
Examples of Errors

Duplicate Structures



Structure	Formula	FW	CAS	NAME	MP	EstMP	ErrorMP
	$C_3H_6O_3$	90.0779	000050-21-5	LACTIC ACID	1.680000000000000e+001	2.266000000000000e+001	5.860000000000000e+000
	$C_3H_6O_3$	90.0779	000079-33-4	L-LACTIC ACID	5.300000000000000e+001	2.266000000000000e+001	-3.034000000000000e+001
	$C_3H_6O_3$	90.0779	000598-82-3	A-HYDROXYPROPIONIC ACID	1.800000000000000e+001	2.266000000000000e+001	4.660000000000000e+000
	$C_3H_6O_3$	90.0779	010326-41-7	D-LACTIC ACID	5.280000000000000e+001	2.266000000000000e+001	-3.014000000000000e+001

Covalent Halogens



Mol Block	CAS	NAME	Smiles
	000056-93-9	BENZYL TRIMETHYL AMMONIUM CHLORIDE	
	000068-05-3	TETRAETHYL AMMONIUM IODIDE	
	000071-91-0	TETRAETHYL AMMONIUM BROMIDE	

Mansouri et al. SAR QSAR Environ. Res. 2016, 27 (11), 939−965.

Summary:

Property	Initial file flagged	Updated 3-4 STAR	Curated QSAR ready
AOP	818	818	745
BCF	685	618	608
BioHC	175	151	150
Biowin	1265	1196	1171
BP	5890	5591	5436
HL	1829	1758	1711
KM	631	548	541
KOA	308	277	270
LogP	15809	14544	14041
MP	10051	9120	8656
PC	788	750	735
VP	3037	2840	2716
WF	5764	5076	4836
WS	2348	2046	2010

Mansouri et al. SAR QSAR Environ. Res. 2016, 27 (11), 939−965.

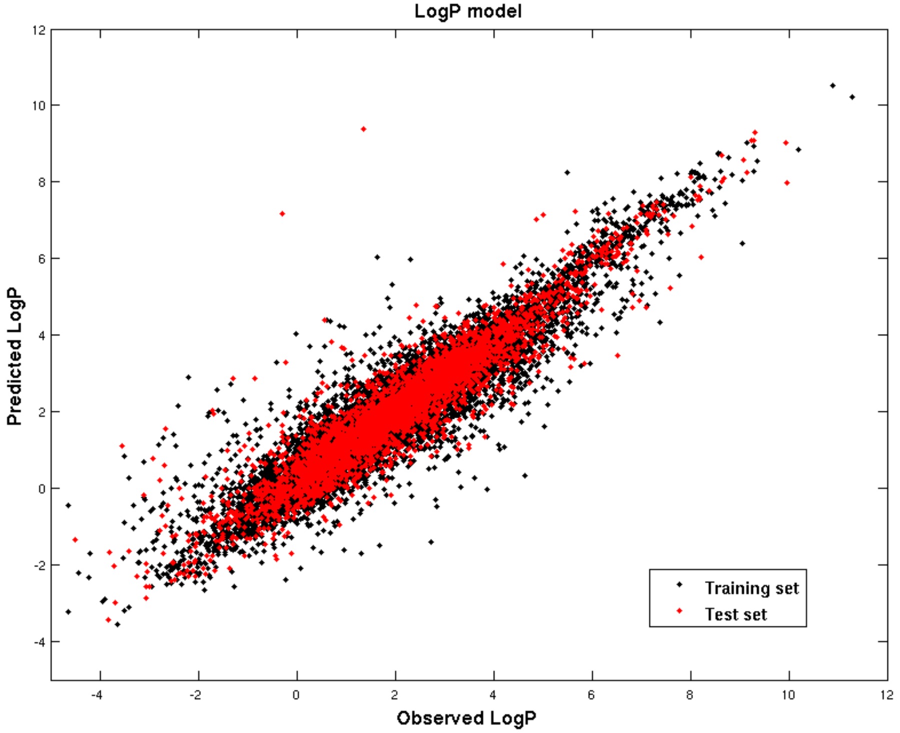
OPERA models

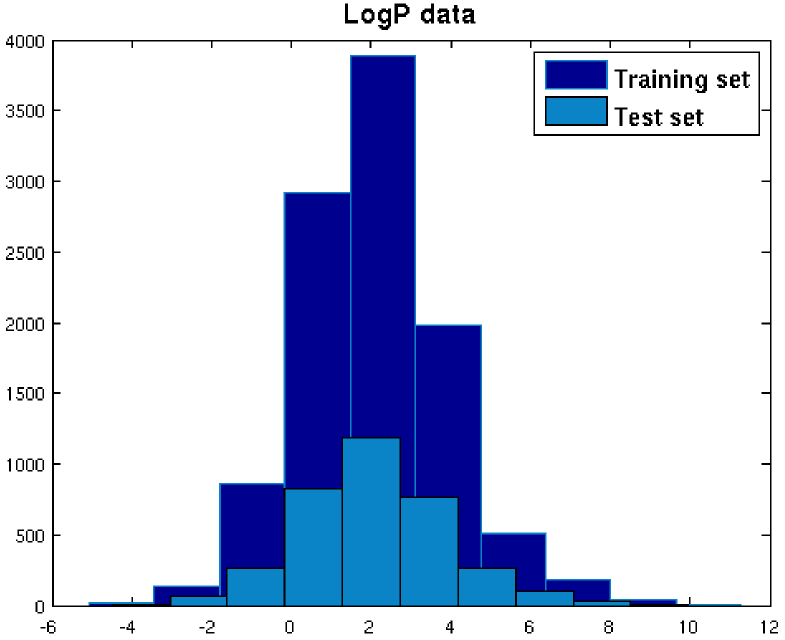
Prop	Vars	5-fold CV (75%)		Training (75%)			Test (25%)		
		Q2	RMSE	N	R2	RMSE	N	R2	RMSE
BCF	10	0.84	0.55	465	0.85	0.53	161	0.83	0.64
BP	13	0.93	22.46	4077	0.93	22.06	1358	0.93	22.08
LogP	9	0.85	0.69	10531	0.86	0.67	3510	0.86	0.78
MP	15	0.72	51.8	6486	0.74	50.27	2167	0.73	52.72
VP	12	0.91	1.08	2034	0.91	1.08	679	0.92	1
WS	11	0.87	0.81	3158	0.87	0.82	1066	0.86	0.86
HL	9	0.84	1.96	441	0.84	1.91	150	0.85	1.82

OPERA models

Prop	Vars	5-fold CV (75%)		Training (75%)			Test (25%)		
		Q2	RMSE	N	R2	RMSE	N	R2	RMSE
AOH	13	0.85	1.14	516	0.85	1.12	176	0.83	1.23
BioHL	6	0.89	0.25	112	0.88	0.26	38	0.75	0.38
KM	12	0.83	0.49	405	0.82	0.5	136	0.73	0.62
KOC	12	0.81	0.55	545	0.81	0.54	184	0.71	0.61
KOA	2	0.95	0.69	202	0.95	0.65	68	0.96	0.68
		BA	Sn-Sp		BA	Sn-Sp		BA	Sn-Sp
R-Bio	10	0.8	0.82-0.78	1198	0.8	0.82-0.79	411	0.79	0.81-0.77

LogP Model: Weighted kNN Model, 9 descriptors





Weighted 5-nearest neighbors

9 Descriptors

Training set: 10531 chemicals

Test set: 3510 chemicals

5 fold Cross-validation:

Q2=0.85 RMSE=0.69

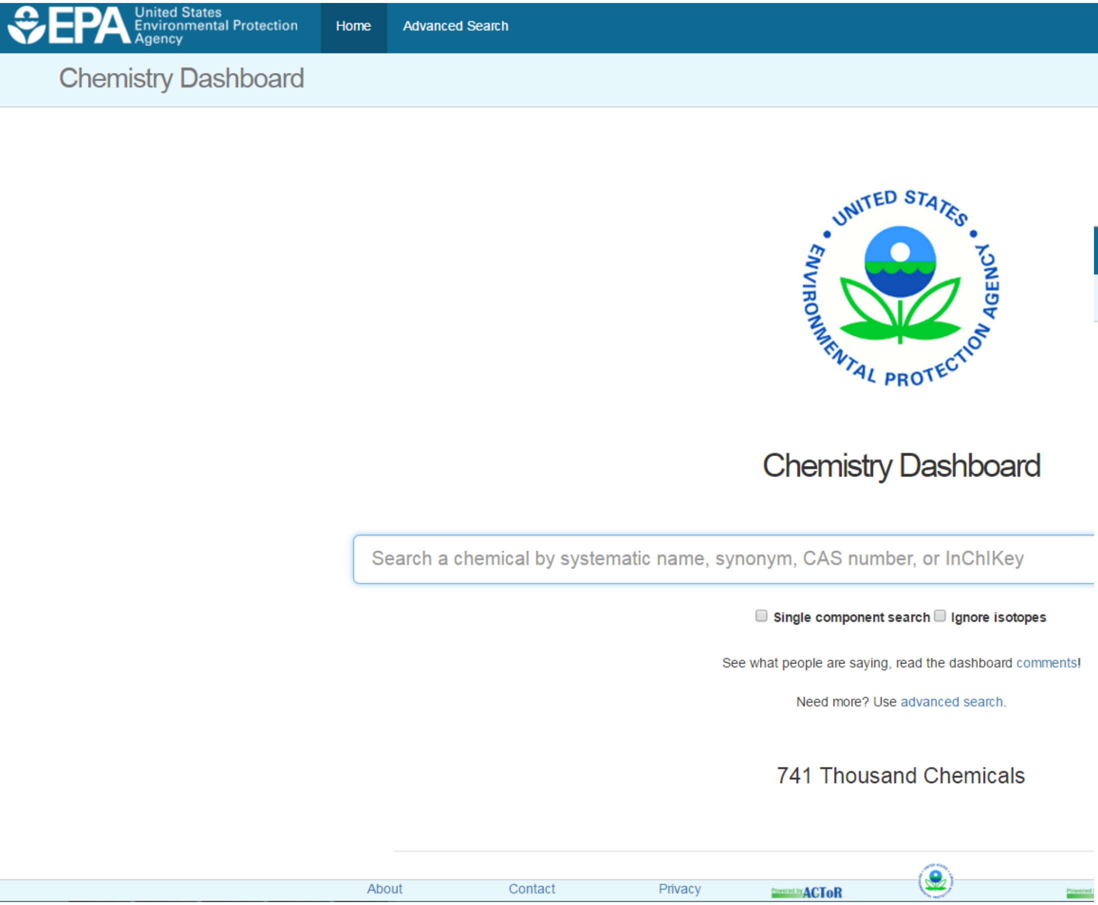
Fitting:

R2=0.86 RMSE=0.67

Test:

R2=0.86 RMSE=0.78

The iCSS Chemistry Dashboard at https://comptox.epa.gov



Chemistry Dashboard

Chemistry Dashboard

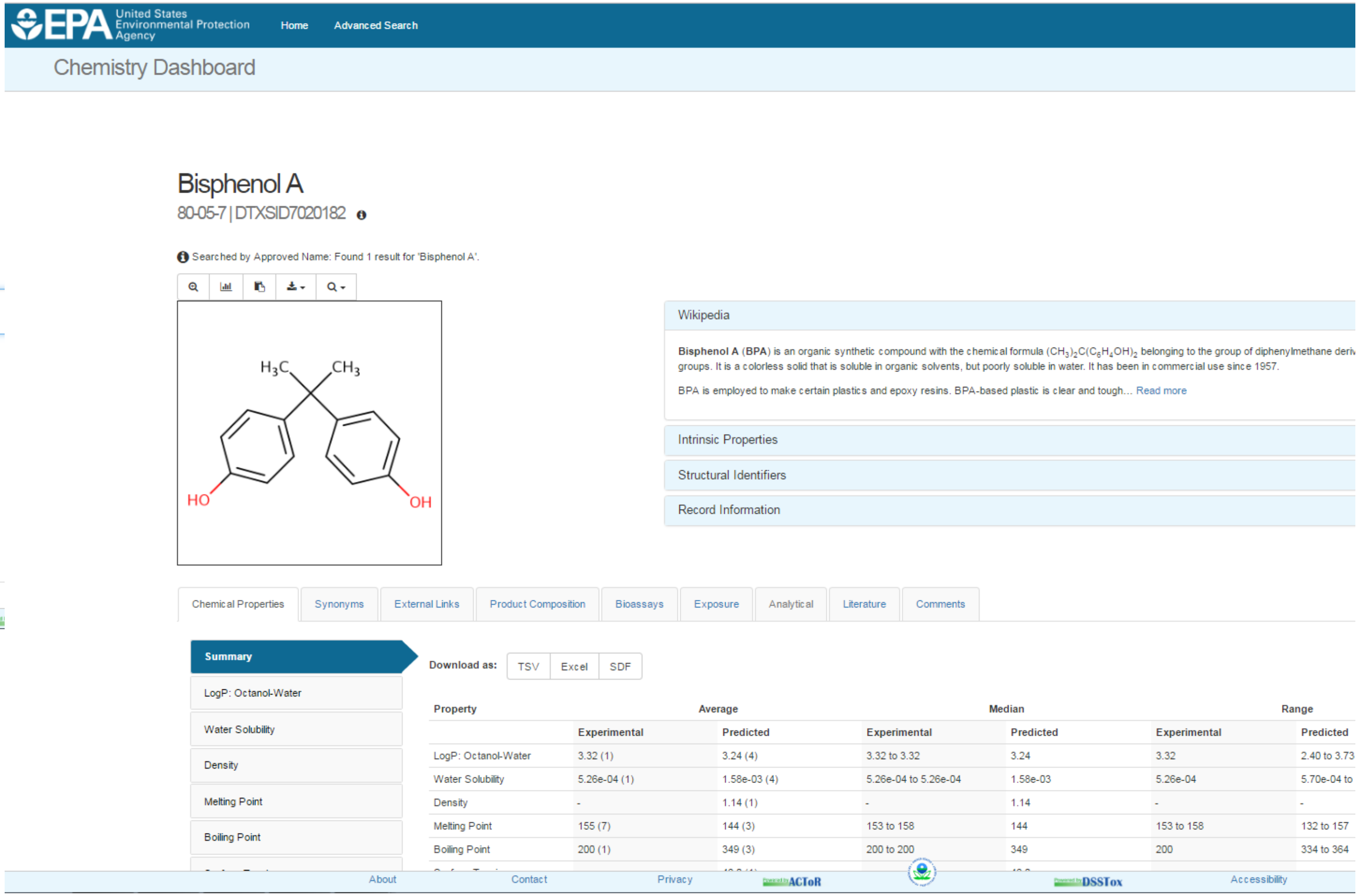
Search a chemical by systematic name, synonym, CAS number, or InChIKey

□ Single component search □ Ignore isotopes

See what people are saying, read the dashboard comments!

Need more? Use advanced search.

741 Thousand Chemicals



Chemistry Dashboard

Bisphenol A

80-05-7 | DTXSID7020182

Searched by Approved Name: Found 1 result for 'Bisphenol A'.

Wikipedia

Bisphenol A (**BPA**) is an organic synthetic compound with the chemical formula $(CH_3)_2C(C_6H_4OH)_2$ belonging to the group of diphenylmethane deriv
groups. It is a colorless solid that is soluble in organic solvents, but poorly soluble in water. It has been in commercial use since 1957.

BPA is employed to make certain plastics and epoxy resins. BPA-based plastic is clear and tough... Read more

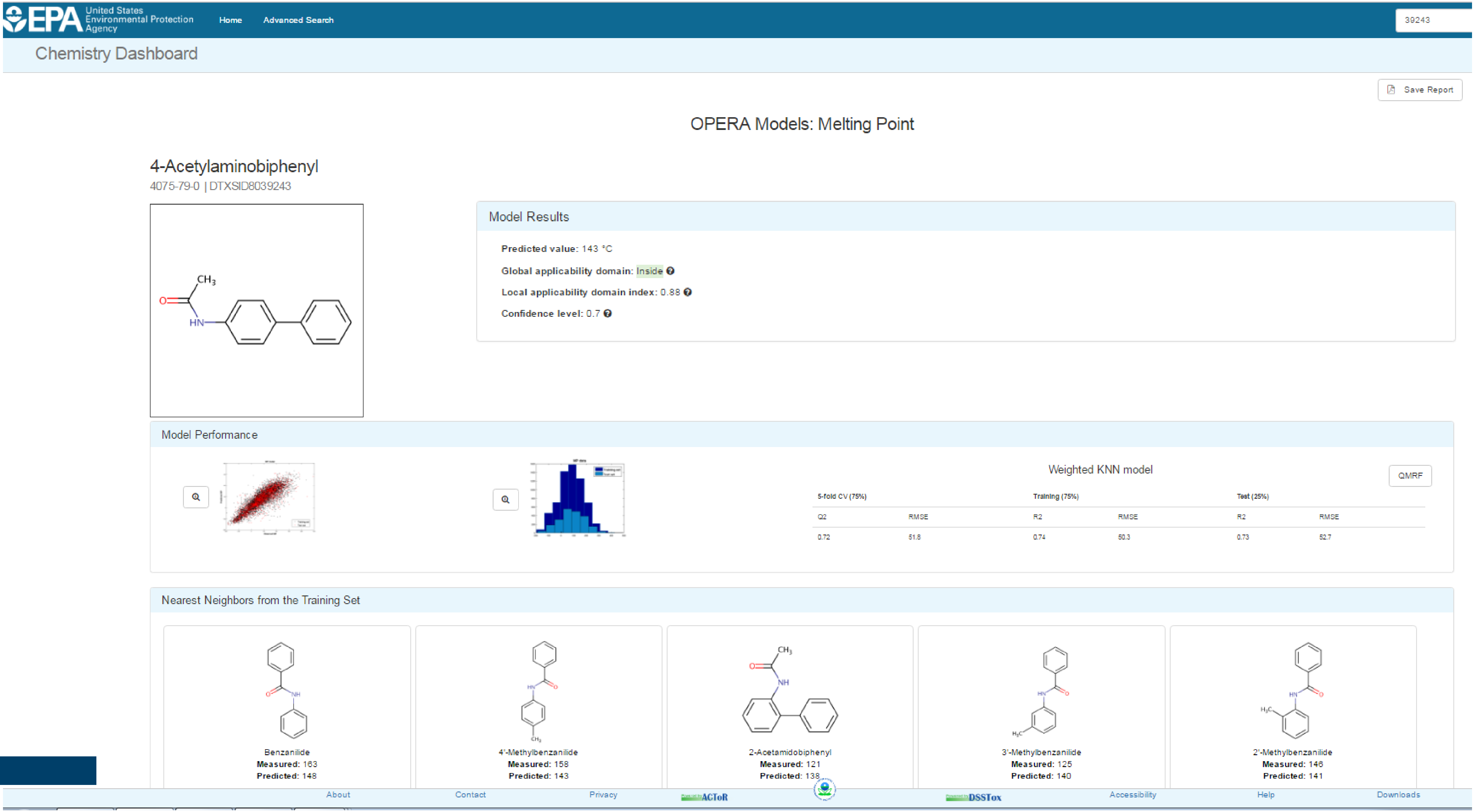
Intrinsic Properties

Structural Identifiers

Record Information

Download as: TSV Excel SDF

Property	Average		Median		Range	
	Experimental	Predicted	Experimental	Predicted	Experimental	Predicted
LogP: Octanol-Water	3.32 (1)	3.24 (4)	3.32 to 3.32	3.24	3.32	2.40 to 3.73
Water Solubility	5.26e-04 (1)	1.58e-03 (4)	5.26e-04 to 5.26e-04	1.58e-03	5.26e-04	5.70e-04 to
Density	-	1.14 (1)	-	1.14	-	-
Melting Point	155 (7)	144 (3)	153 to 158	144	153 to 158	132 to 157
Boiling Point	200 (1)	349 (3)	200 to 200	349	200	334 to 364



OPERA Models: Melting Point

4-Acetylaminobiphenyl

4075-79-0 | DTXSID8039243

Model Results

Predicted value: 143 °C

Global applicability domain: Inside

Local applicability domain index: 0.88

Confidence level: 0.7

Model Performance

Weighted KNN model

5-fold CV (75%)		Training (75%)		Test (25%)	
Q2	RMSE	R2	RMSE	R2	RMSE
0.72	51.8	0.74	50.3	0.73	52.7

Nearest Neighbors from the Training Set

Benzanilide	4'-Methylbenzanilide	2-Acetamidobiphenyl	3'-Methylbenzanilide	2'-Methylbenzanilide
Measured: 163	**Measured:** 158	**Measured:** 121	**Measured:** 125	**Measured:** 146
Predicted: 148	**Predicted:** 143	**Predicted:** 138	**Predicted:** 140	**Predicted:** 141

Acknowledgements

National Center for Computational Toxicology

United States
Environmental Protection
Agency

Thank you for your attention