



Welcome !

This Session will Begin at 2:00 pm Eastern US Time

**“Experiments are too hard: How to use online resources
for predictive toxicology”**

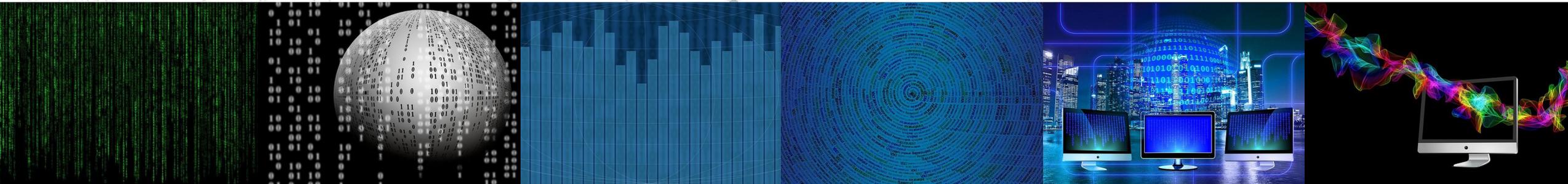
Sciome LLC

Ruchir Shah

Eric McAfee

Alex Sedykh

Vijay Gombar Austin Ross

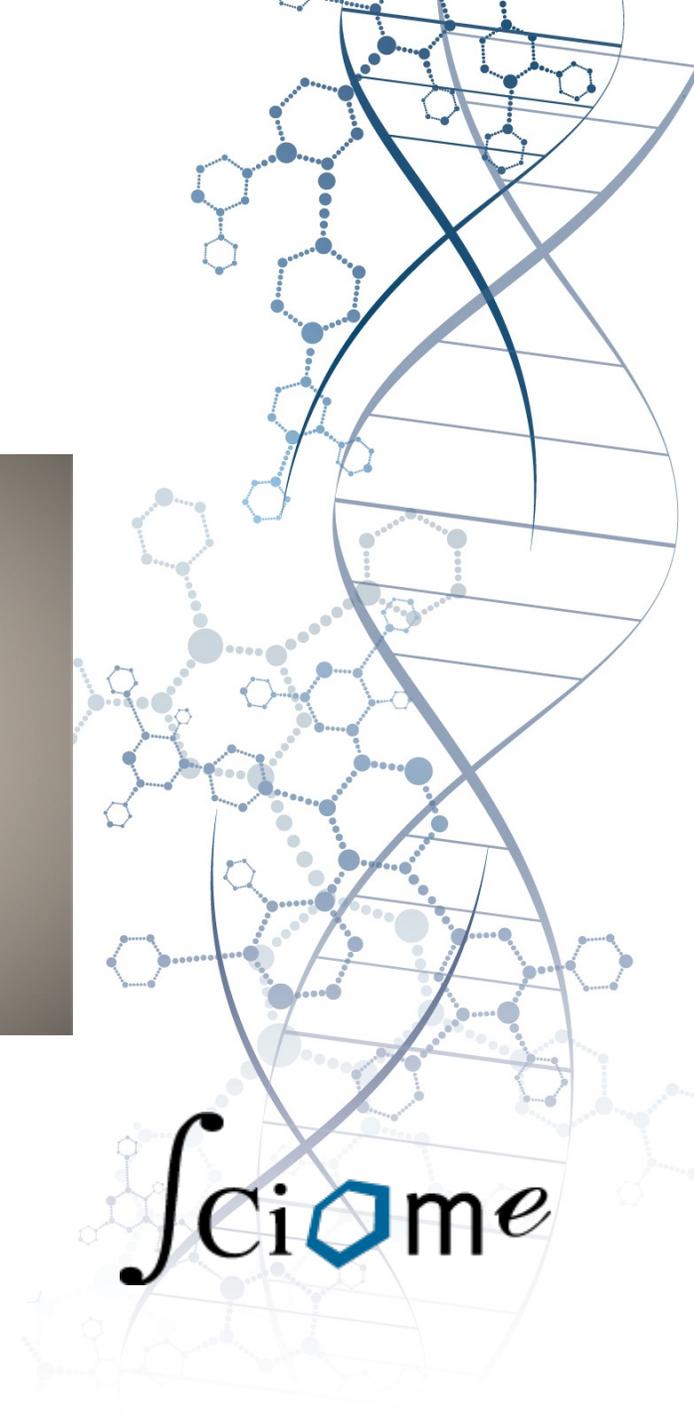


Ruchir Shah, Ph.D.

Chief Scientific Officer

ruchir.shah@sciome.com

www.sciome.com



BIG DATA IN ENVIRONMENTAL SCIENCE AND TOXICOLOGY

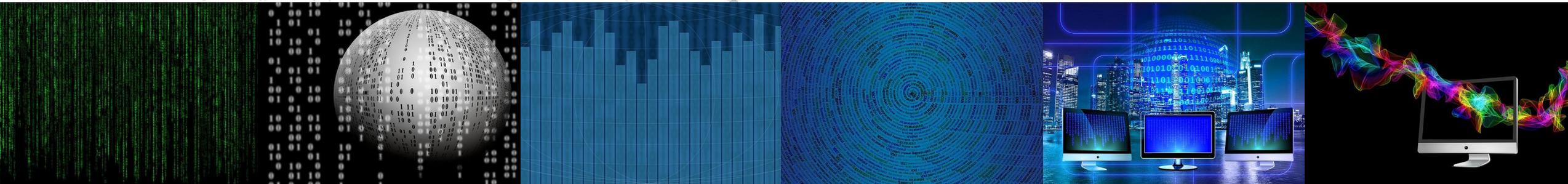
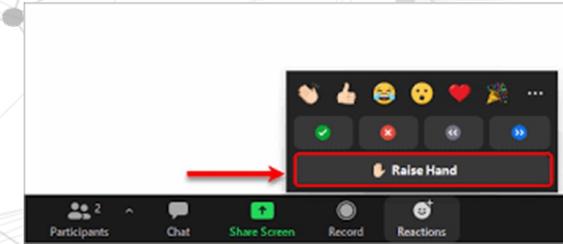


superfund.tamu.edu



TEXAS A&M UNIVERSITY
Superfund Research Center

- **All participants are muted to enable the speaker to present without interruption.**
- **Please rename yourself and designate Full Name and Affiliation.**
- **Use the Chat or Reaction icon at the bottom of your screen.**
- **This meeting will be recorded, and posted on the @tamusuperfund website <https://superfund.tamu.edu/big-data-series-2021/> in the coming weeks.**



Bioinformatics

- ✓ Next-Generation Sequence data analysis
- ✓ Microarray data analysis
- ✓ Structural & Functional genomics
- ✓ SNP/Genotype analysis & GWAS
- ✓ Biostatistics and Mathematical Modeling

Cheminformatics

- ✓ Quantitative Structure-Activity Relationship (QSAR) modeling
- ✓ Computational Toxicity Predictions
- ✓ Active site and Protein-Protein Docking
- ✓ Pharmacophore Modeling

Text-Mining and Literature Review

- ✓ Document Tagging and Visualization
- ✓ Full-Text Conversion and Search
- ✓ Document Clustering, Ranking & Classification
- ✓ Literature Prioritization and Screening
- ✓ Data extraction
- ✓ rapid Evidence Mapping (rEM) and systematic reviews
- ✓ Web mining and information retrieval

Data Science and Analytics

- ✓ Integration and visualization of large volumes of heterogeneous data
- ✓ Development and implementation of Deep Learning methodologies for predictive science
- ✓ Automated Image analysis using artificial intelligence
- ✓ Natural Language Processing (NLP) methods using Deep Learning

Software Development

- ✓ Requirements gathering
- ✓ Software architecture design
- ✓ User interface design
- ✓ Implementation, deployment and maintenance
- ✓ User support

25 Full time Informaticians

>Half with PhD, Most with a Masters

All of us program, develop methods, analyzed data, and publish

~190 total publications, 2 patents

Sciome, Fall 2021



Open Positions at Sciome

- **Bioinformatics Scientist**
- **Cheminformatics Scientist**
- **Data Scientist / ML Engineer / NLP expert**
- **Software Developer**
- **Statistician**

Publicly available predictive models and resources

OCHEM: <https://ochem.eu/home/show.do>

Contains 3917726 records for 923 properties (with at least 50 records) collected from 16996 sources

Available models: LogP, LogS, Solubility in DMSO, Melting point, Boiling point, CYP 450 Inhibition, AhR Activation, AMES Test, BioConcentration factor, T. Pyriformis toxicity, Bioavailability, Gastrointestinal absorption, BBB permeability, CACO-2

OECD QSAR Toolbox: <https://qsartoolbox.org/>

The Toolbox is a software application intended to be used by governments, chemical industry and other stakeholders in filling gaps in (eco)toxicity data needed for assessing the hazards of chemicals. The Toolbox incorporates information and tools from various sources into a logical workflow.

59 databases containing ~100 000 chemicals with above 3 million measured data points (<https://qsartoolbox.org/resources/databases/>)

ECOSAR: <https://www.epa.gov/tsca-screening-tools/ecological-structure-activity-relationships-ecosar-predictive-model>

A computerized predictive system that estimates aquatic toxicity. The program estimates a chemical's acute (short-term) toxicity and chronic (long-term or delayed) toxicity to aquatic organisms, such as fish, aquatic invertebrates, and aquatic plants, by using computerized Structure Activity Relationships (SARs).

EPI Suite: <https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface>

A suite of physical/chemical property and environmental fate estimation programs developed by EPA's and Syracuse Research Corp. (SRC). Includes ECOSAR.

ICE: <https://ice.ntp.niehs.nih.gov/>

Provides curated data from NICEATM (National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM)), its partners, and other resources, as well as tools to facilitate the safety assessment of chemicals.

OPERA: <https://ntp.niehs.nih.gov/whatwestudy/niceatm/comptox/ct-opera/opera.html>

To provide robust QSAR/QSPR models for chemical properties of environmental interest that can be used for regulatory purposes

Overview

Mr. Eric McAfee: .. Will present NTP's ICE tool, which hosts curated data from NICEATM and its partners.

Dr. Vijay Gombar: .. Will present the OPERA tool followed by latest research in predictive toxicology using SAAGAR features and OrbiTox

Mr. Austin Ross: .. Will demonstrate OrbiTox

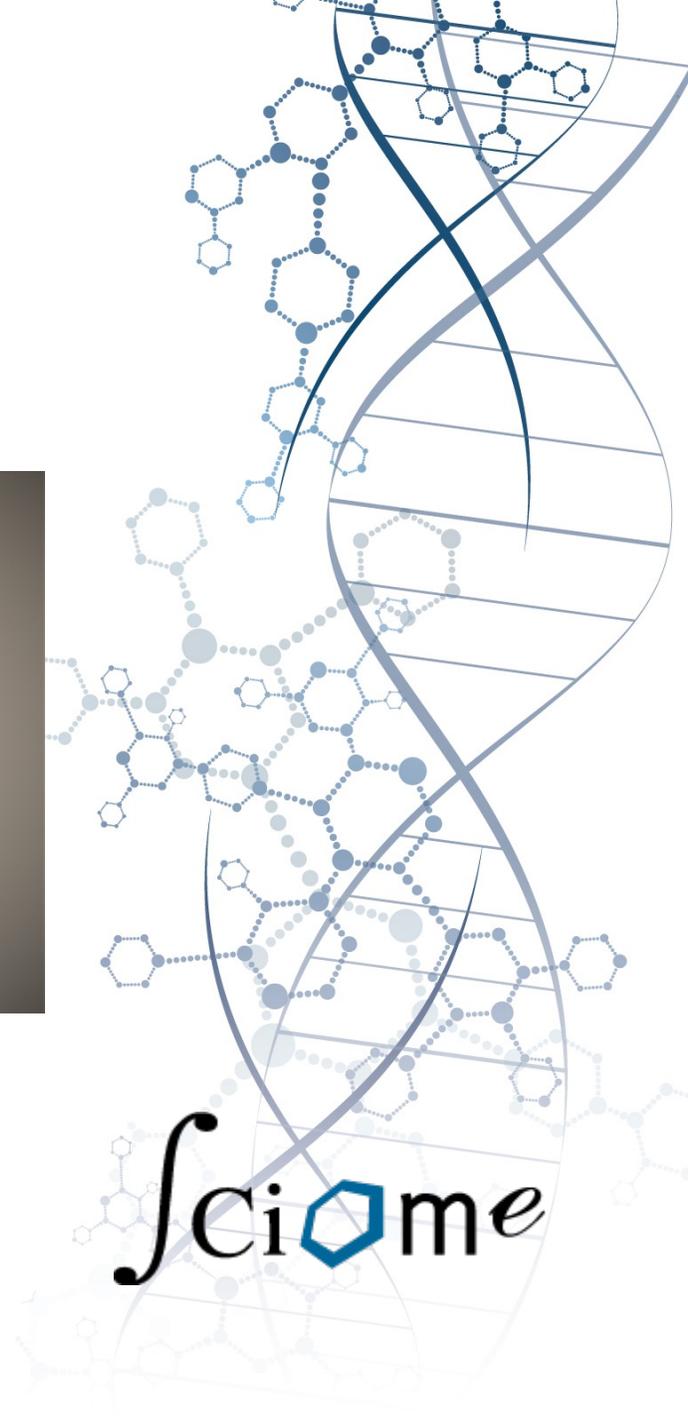
Eric McAfee

Sr. Software Developer

Eric.McAfee@sciome.com



Sciome





ICE (Integrated Chemical Environment)

Sciome: Ruchir Shah, Eric McAfee, Alex Sedykh, Vijay Gombar, Austin Ross

12/01/2021



Sciome

What is ICE?

<https://ice.ntp.niehs.nih.gov/>

- Portal with high-quality data for computational workflows
- Data curated and aggregated from various NTP sources:
 - NICEATM (NTP Interagency Center for the Evaluation of Alternative Toxicological Methods)
 - EPA (Environmental Protection Agency)
 - CEBS (Chemical Effects in Biological Systems)
- Support for inputs of multiple chemical identifiers:
 - CASRNs
 - DTXSIDs
 - SMILES
 - InChIKeys (International Chemical Key)
- Chemical Lists – Curated lists, known activities
- Assay Groups – In Vitro, In Vivo, In Silico
- Dose response data, find adverse reactions (AC50, LD50)

The screenshot displays the ICE website interface. At the top, there is a navigation bar for the National Toxicology Program (U.S. Department of Health and Human Services) and the Integrated Chemical Environment. A search bar is present on the right. Below the navigation bar, the main content area features a 'News & Events' section with a 'ICE v3.5 Release' announcement, listing new tools and capabilities such as Saagar Fingerprints (Beta), Individually Selected Results - Curve Surfer and Chemical Quest, SDF Downloads, and AC50 Plots. A 'Learn about ICE updates' button is provided. To the right, there is a promotional banner for 'Read the latest on ICE' featuring a book cover titled 'COMPUTATIONAL TOXICOLOGY' and a 'Recent ICE publication of tools that support chemical evaluations' article by Abedini et al. Below these sections, a grid of seven icons represents various tools: Search, Chemical Quest, Curve Surfer, PBPK, IVIVE, Chemical Characterization, and Data. At the bottom, there is a 'BACK TO TOP' link and a footer note: 'Web page last updated on Feb. 21, 2020'.

Chemical Search

- Support for individual chemicals and mixtures
 - Support for inputs of multiple chemical identifiers
 - Integrated view
 - Data Downloads
 - Text
 - Excel
 - PDF
 - SDF
 - Detail Visualizations
 - Chemical
 - Mixture
 - Summary Visualizations
 - Activity Call Data
 - AC50 values
 - By Chemical
 - By Assay
 - Downloads

Integrated Chemical Environment

HOME SEARCH TOOLS DATA ABOUT HELP

Input Results Search Results

Acetoc

Active AC50 endpoints for Steroid Hormone Metabolism by assay (94 assays)

Select Assay

Active

Finished

Active AC50 endpoints for Steroid Hormone Metabolism by assay (94 assays)

Download icons: TXT, XLSX

Assay	Chemical Name	CASRN (CEBS Link)	DTXSID (Dashboard Link)	Value
ACEA_AR_agonist_80hr	17beta-Trenbolone	CASRN: 10161-33-8	DTXSID: DTXSID0034192	3.99197174... 4
ACEA_AR_agonist_80hr	Spironolactone	CASRN: 52-01-7	DTXSID: DTXSID6034186	0.04462739...
ACEA_AR_agonist_80hr	5alpha-Dihydrotestosterone	CASRN: 521-18-6	DTXSID: DTXSID9022364	1.73029467... 4
ACEA_AR_agonist_80hr	17alpha-Estradiol	CASRN: 57-91-0	DTXSID: DTXSID8022377	0.00843748...
ACEA_AR_agonist_80hr	17-Methyltestosterone	CASRN: 58-18-4	DTXSID: DTXSID1033664	5.65914093...

Close

Close

Tools and Workflows

- Curve Surfer
 - The Curve Surfer tool allows the user to view and interact with concentration response curves from cHTS
- PBPK (Physiological based pharmacokinetic modeling and simulation)
 - PBPK tool allows the user to generate predictions of tissue-specific chemical concentration profiles following a dosing event
- IVIVE (In Vitro to In Vivo Extrapolation)
 - The IVIVE tool uses pharmacokinetic models to predict the equivalent administered dose (EAD) from the activity concentration of selected assays
- Chemical Characterization
 - The Chemical Characterization tool allows the user to view and compare one or two chemical lists based on their physicochemical properties
- Chemical Quest
 - The Chemical Quest tool uses fingerprints to predict structure similarity

Curve Surfer

- Curve Surfer - view and interact with concentration response curves from cHTS
- Input a chemical list and select assays
- Output Results
 - Downloaders
 - Text
 - Excel
 - PDF
 - Filter chain
 - Dose response curves (overlay coming)

The screenshot displays the Curve Surfer web application interface. At the top, there is a navigation bar with 'HOME', 'SEARCH', 'TOOLS', 'DATA', 'ABOUT', and 'HELP'. Below this, a secondary navigation bar includes 'Chemical Quest', 'Curve Surfer' (highlighted), 'PBPK', 'IVIVE', and 'Chemical Characterization'. The main interface features a 'Send filtered results to:' section with options for TXT, XLSX, and PDF, and a 'Clear Filter' button. Below this, there are controls for 'Select Page' (1 of 6), 'Showing 1-10 of 58 curves', 'Sort Results By' (Chemical Name), and 'Direction' (Asc). There are also filters for 'Select Mechanistic Target To View Curves' (All), 'Assay Text Filter', 'Select Assay(s)' (0 values), 'Select CASRN(s)' (0 values), and 'Select Call(s)' (Active). A 'Select All Filtered' button and 'Clear Selected' button are present, along with a checkbox for 'Only show selected items' and a 'Selected Item(s): 0/156' indicator.

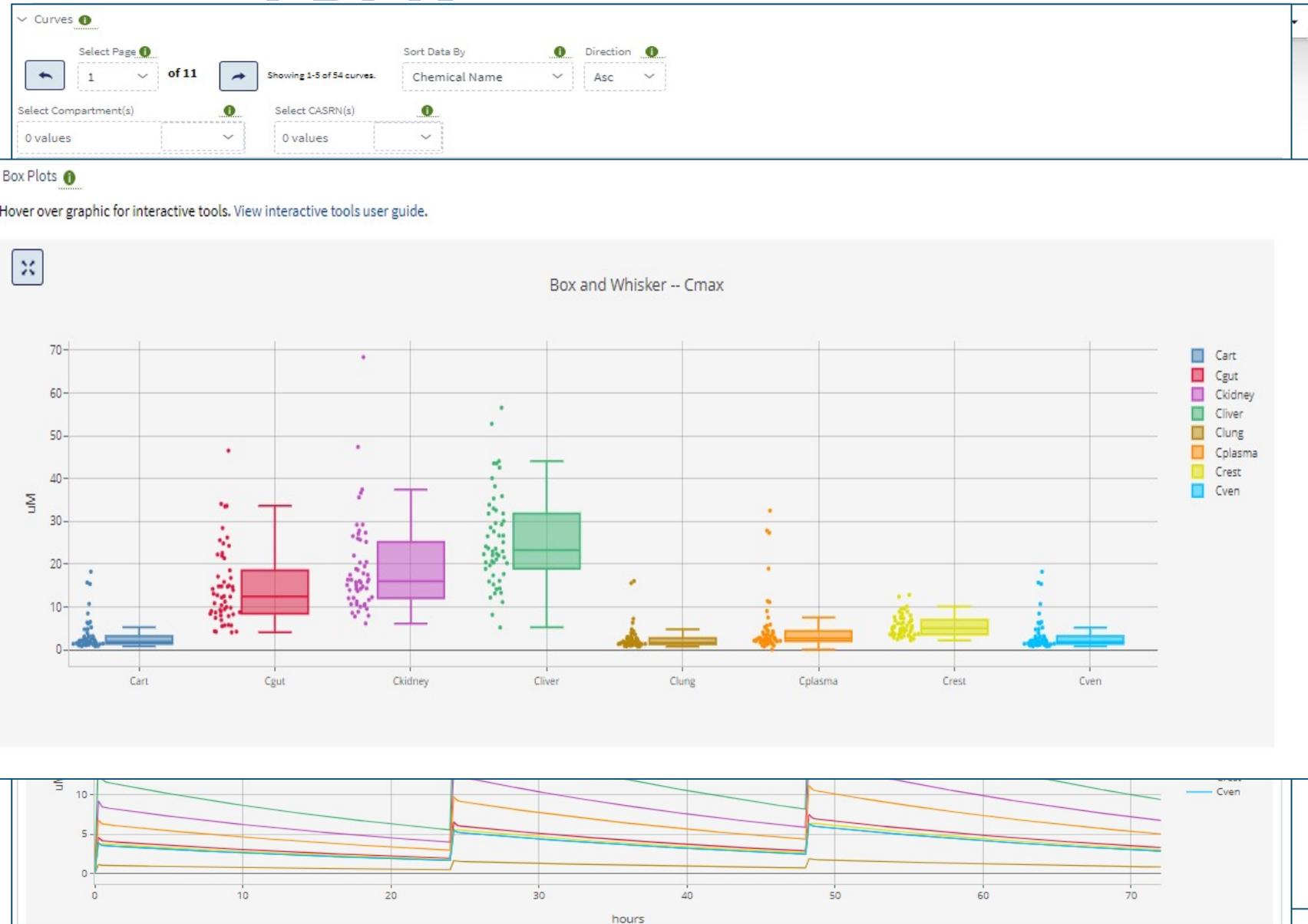
Two assay cards are shown below the filter chain:

- Assay: BSK_LPS_PGE2_up**
CASRN: 58-18-4
Chemical Name: 17-Methyltestosterone
LOEC (uM): 40.0
Call: Active
Mechanistic Target: Inflammatory Response
DTXSID: DTXSID1033664
Winning Curve-Fit Model: Hill
Top of Curve: 0.23
- Assay: TOX21_RT_HEK293_FLO_08hr_viability**
CASRN: 58-18-4
Chemical Name: 17-Methyltestosterone
AC50 (uM): 2.9
Top of Curve: 27.49
Mechanistic Target: Cell Viability Process
DTXSID: DTXSID1033664
Winning Curve-Fit Model: Hill
ACC (uM): 3.63
Call: Active

Each assay card includes a plot showing the concentration response curve. The left plot for BSK_LPS_PGE2_up shows log10 fold induction vs Concentration (uM) on a log scale. The right plot for TOX21_RT_HEK293_FLO_08hr_viability shows percent activity vs Concentration (uM) on a log scale. Both plots include a Hill fit curve, concentration response data points, and horizontal dashed lines for LOEC, AC50, and Top of Curve.

PBPK

- PBPK (Physiological based pharmacokinetic modeling and simulation)
 - PBPK tool allows the user to generate predictions of tissue-specific chemical concentration profiles following a dosing event
- Input a chemical list and PBPK parameters
- Output
 - Download files
 - PBPK Results
 - Chemical
 - Compartment
 - CSS
 - Cmax
 - PBPK Results Visualizations
 - Does response curves with compartment overlay
 - Box Plots (Cmax across compartments)



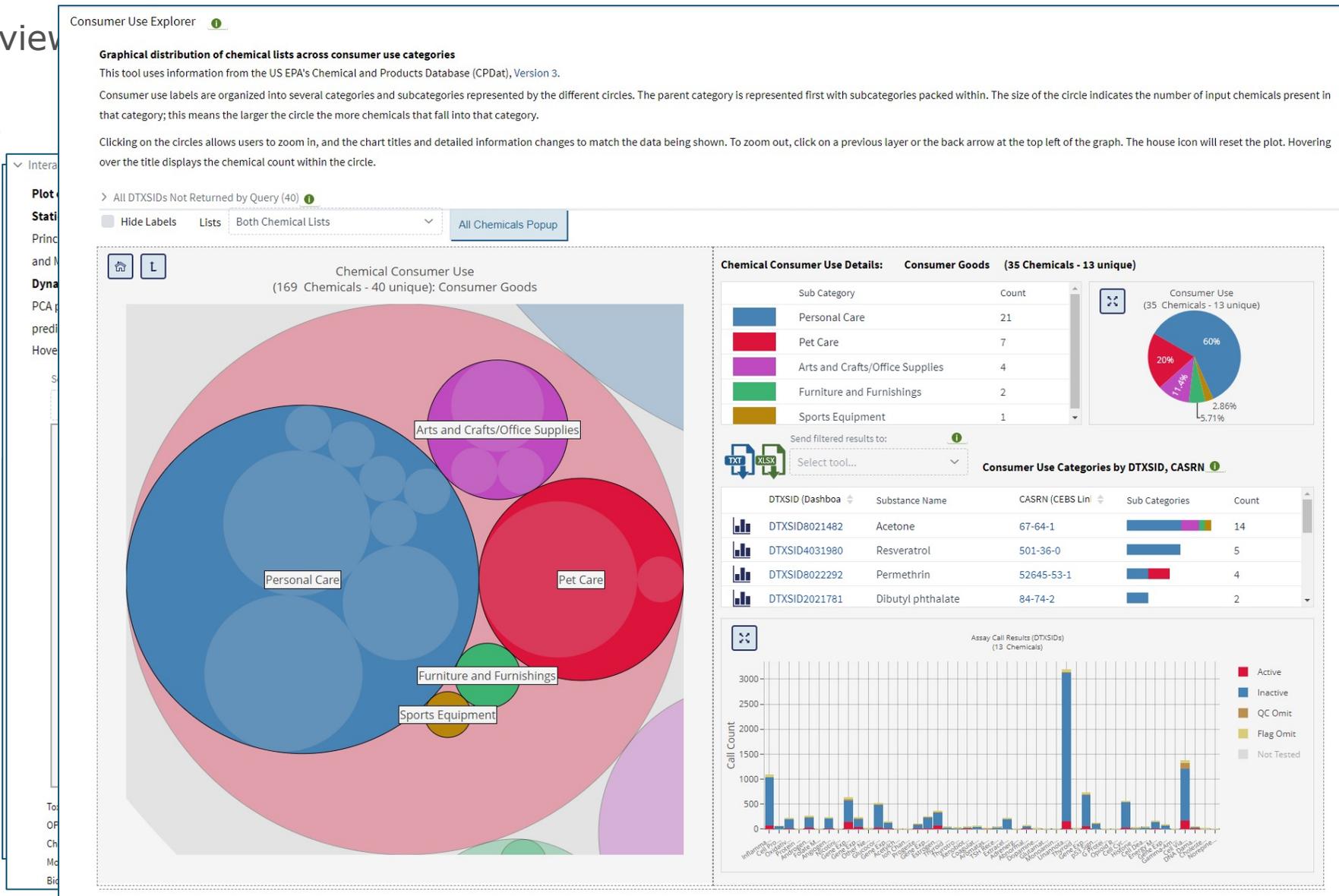
IVIVE

- IVIVE (In Vitro to In Vivo Extrapolation) - uses pharmacokinetic models to predict the equivalent administered dose (EAD) from the activity concentration of selected assays
- Input a chemical list, select assays, params
- Output
 - Download files
 - TXT
 - XLSX
 - IVIVE Results
 - Chemical
 - Assay
 - Mode of Action
 - Mechanistic Targets
 - AC50
 - EAD 50%
 - EAD 95%
 - IVIVE Results Visualizations
 - EAD 95th Box and Whisker
 - Distribution of in vitro bioactivity (AC50)



Chemical Characterization

- Chemical Characterization - view and compare one or two chemical lists based on their physicochemical properties
- Input 2 different lists for comparison
- Output
 - Chemical Properties Summary
 - Visualization of Chemical Properties
 - Interactive PCA
 - Consumer Use Explorer
 - Zoom in



Chemical Quest

- Chemical Quest tool uses fingerprints to predict structure similarity
- Input a chemical list and list of structures

- Draw structures
- add smiles to input list

- Output Results

- Tanimoto threshold
- Saagar
- Downloaders
 - Text
 - Excel
 - SDF
- Chemical Structures
- Filter Chain

Similar Structures to: CON(C)C(=O)NC1=CC(Cl)=C(Br)C=C1

Select Page: 1 of 1 | Showing 1-10 of 10 hits | Sort Results By: Tanimoto | Direction: Desc

Tanimoto Filter: [] | SMARTS Filter: []

Select All Filtered | Clear Selected | Only show selected items

CASRN	DTXSID	Chemical Name	Tanimoto Value
13360-45-7	DTXSID6040290	Chlorbromuron	1.0
330-55-2	DTXSID2024163	Linuron	0.956522
NOCAS_892686	DTXSID20892686	N'-(3,4-Dichlorophenyl)-N-methoxy-N-methyl(-2-H)urea	0.956522
60095-90-1	DTXSID20608603	N'-(2,3-Dichlorophenyl)-N-methoxy-N-methylurea	0.888889
17356-61-5	DTXSID70864772	N'-(3,4-Dichlorophenyl)-N-methoxyurea	0.888889
1746-81-2	DTXSID0037576	Monolinuron	0.873239
3060-89-7	DTXSID6042157	Metobromuron	0.871429
1630-19-9	DTXSID60540391	N'-(3-Chlorophenyl)-N-methoxy-N-methylurea	0.871429
102636-55-5	DTXSID90761214	N'-(4-Chloro-3-hydroxyphenyl)-N-methoxy-N-methylurea	0.842105
149282-25-7	DTXSID50375699	N-[(4-Chlorophenyl)methyl]-N'-(3,4-dichlorophenyl)-N-methoxyurea	0.825

Vijay K Gombar, Ph.D.

Cheminformatics Scientist

Vijay.Gombar@sciome.com



OPERA – OPEn (q)saR App: Introduction

Credits to: Dr. Kamel Mansouri

Computational Chemist at National Institute of Environmental Health Sciences (NIEHS), RTP, NC

- Developed by US (National Center for Computational Toxicology (Mansouri et al. 2018)¹.
- OPERA can be downloaded from the National Institute of Environmental Health Sciences GitHub repository (<https://github.com/NIEHS/OPERA>) or <https://github.com/kmansouri/OPERA>

Both a command-line version and user-friendly graphical user interface versions are available for Windows and Linux operating systems

¹Mansouri K, Grulke CM, Judson RS, Williams AJ (2018). OPERA models for predicting physicochemical properties and environmental fate endpoints. J Cheminform 10(1):10, PMID: 29520515, <https://doi.org/10.1186/s13321-018-0263-1>.

OPERA – OPEN (q)saR App: Interface

The screenshot displays the OPERA 2.3 application window. The interface includes several sections:

- Input:** A text field with an information icon and a "Browse" button.
- Output:** A text field with an information icon and a "Browse" button.
- Models:** A list of checkboxes for various models: Physchem (with sub-options LogP and LogBCF), Environme, Toxicity en (with sub-option ER (CE)), ADME prop, and FUB.
- Output options:** A list of checkboxes: Separate files, Experimental values, Nearest neighbors, Include descriptor values, and Keep full descriptors files.
- Results summary:** A large empty text area with an information icon.
- Buttons:** "Standardize" and "Calculate" buttons are visible on the right side.

A dialog box titled "Output file" is open, providing information about output file formats:

By default, the output file(s) will contain the predictions, applicability domain and accuracy assessment.
The output file extension could be:

- .csv:** comma "," delimited csv file with headers. One molecule/row. Provided or generic ID in first column.
- .txt:** Text file with multiple rows/structure. Can be used as a prediction report.

An "OK" button is located at the bottom of the dialog box.

OPERA – OPEn (q)saR App: Models

Input i C:\OPERA\Test\Example1.smi

Output i C:\OPERA\Test\Example1_OPERA2.3_Pred.csv

Models i

Physchem properties
 LogP MP BP VP

Environmental fate
 LogBCF AOH Bio

Toxicity endpoints
 ER (CERAPP) AR ((

ADME properties
 FUB Clint

Output options i

Separate files
 Experimental values
 Nearest neighbors
 Include descriptor values
 Keep full descriptors files

Physicochemical properties:

LogP v2.0	OC	CC(=O)OO	DTXSID1025853
MP v1.5	Me	CCOP(=S)(OCC)Oc1ccc([N+](=O)[O-])cc1	DTXSID7021100
BP v1.5	Bo	COP(=O)(OC)OC=C(Cl)Cl	DTXSID5020449
VP v1.5	Va	Oc1c(Cl)cc(Cl)c(Cl)c1Cc1c(O)c(Cl)cc(Cl)c1Cl	DTXSID6020690
WS v2.1	Wa	OCCOc1ccccc1	DTXSID9021976
		BrCCBr	DTXSID3020415
		c1ccccc1	DTXSID3039242
		(Cl)c(Cl)c(Cl)c(C#N)c1Cl	DTXSID0020319
		c(O)c1	DTXSID2021238
		C(C=C(Cl)Cl)ClC(=O)OCc1cccc(Oc2ccccc2)c1	DTXSID8022292

Output options

By default, the output file(s) will contain the predictions, applicability domain and accuracy assessment in csv or txt format.

The default fields are: Molecule ID, predicted value (pred), Applicability domain (AD), Similarity index (Sim_index) and accuracy estimate (Conf_index).

Additional options:

Separate files: Separate output file for each endpoint. Recommended if high number of molecules are predicted.

Experimental values Include the experimental values based on provided CASRNs or DTXSID in the input file.

Nearest neighbors: Includes the (3 or 5) nearest neighbors from training set (CAS, InChIKeys, Observed and predicted values).

Descriptors values: Output file containing all prediction details and used descriptors (only if output is in csv format).

Descriptors files: Keep temporary descriptors files (generated during descriptor calculation).

Models: (acid) dissociation constant, structural properties (StrP v2.0) included by default, MolWeight, nbAtoms, nbHeavyAtoms, nbC, nbO, nbN, nbAromAtom, nbRing, HeteroRing, Sp3Sp2HybRatio, nbRotBd, nbHBdAcc, ndHBdDon, LipinskiFailures, TopoPolSurfAir, MolarRefract, CombDipolPolarizability, absorption half-life in days, biodegradability of organic chemicals, biotransformation primary biotransformation rate constant, distribution coefficient of organic compounds, Androgen Receptor Activity Prediction Project, Estrogen Receptor Antagonist ER activity, Modeling Project for Androgen Receptor Activity, Androgen Receptor Antagonist and Antagonist AR activity, Acute Toxicity Modeling Suite (CATMoS), GHS categories, LD50 (Log mg/kg), plasma fraction unbound, hepatic intrinsic clearance

OPERA – OPEN (q)saR App: Models

1. Physicochemical properties such as acid dissociation constant (pKa) and octanol-water dissociation coefficient (logD) and partition coefficient (logP), water solubility, melting and boiling point (Mansouri et al. 2019)².
2. Ecotoxicity parameters such as fish bioconcentration factor, soil adsorption coefficient, and biodegradability.
3. Parameters for inputs into pharmacokinetic models, such as hepatic clearance and plasma fraction unbound.
4. Estrogenic activity from the Collaborative Estrogen Receptor Activity Prediction Project (CERAPP) (Mansouri et al. 2016)³.
5. Androgenic activity from the Collaborative Modeling Project for Androgen Receptor Activity (CoMPARA) (Mansouri et al. 2020)⁴.
6. Acute oral systemic toxicity from the Collaborative Acute Toxicity Modeling Suite (CATMoS) (Mansouri et al. 2021)⁵.

All OPERA models were built on curated data and QSAR-ready chemical structures standardized using an open-source workflow (Mansouri et al. 2016)⁶.

²Mansouri *et al.* Open-source QSAR models for pKa prediction using multiple machine learning approaches. *J Cheminform* (2019) 11:60, <https://doi.org/10.1186/s13321-019-0384-1>

³Mansouri, el (2016) CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. *Environ. Health Perspect.* 2016, 124 (7), 1023–1033.

⁴Mansouri, et al (2020) CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity. *Environmental Health Perspectives*, 128 (2). CID: 027002, <https://doi.org/10.1289/EHP5580>

⁵Mansouri et al. (2021) CATMoS: Collaborative Acute Toxicity Modeling Suite. *Environmental Health Perspectives*, 129 (4). CID: 047013, <https://doi.org/10.1289/EHP8495>

⁶Mansouri et al (2016) An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSARmodelling. *SARQSAREnvironRes*27(11):911–937, PMID: 27885862, <https://doi.org/10.1080/1062936X.2016.1253611>.

OPERA – OPEn (q)saR App: Models' Relevance

Properties like partition coefficient, boiling point, vapor pressure, and melting point also have great environmental impact.

LogP (Partition coefficient):

Is an important physical property of a substance and, thereby, a predictor of its behavior in different environments.

Is the first indicator on whether a substance will be absorbed by plants, animals, humans, or other living tissue; or be easily carried away and disseminated by water.

Vapor pressure:

Affects a chemical's residence time in soil and in water and is a significant factor in its distribution and transport in the environment.

Melting point/boiling point:

Indicate the physical state of the chemical at ambient temperatures, which will dictate how the chemical is handled and treated.

Kamel Mansouri, Chris M. Grulke, Richard S. Judson & Antony J. Williams. OPERA models for predicting physicochemical properties and environmental fate endpoints. *Journal of Cheminformatics* volume 10, Article number: 10 (2018)

OPERA – OPEn (q)saR App: Running

OPERA 2.3

Input *i* C:\OPERA\Test\Example1.smi

Output *i* C:\OPERA\Test\Example1_OPERA2.3_Pred.csv

Models *i*

Physchem properties
 LogP MP BP VP WS HL KOA RT pKa LogD

Environmental fate
 LogBCF AOH Biodeg R-Biodeg KM KOC

Toxicity endpoints
 ER (CERAPP) AR (CoMPARA) AcuteTox (CATMoS)

ADME properties
 FUB Clint

Output options *i*

Separate files
 Experimental values
 Nearest neighbors
 Include descriptor values
 Keep full descriptors files

Results summary *i*

Loaded structures from SMILES file: 10
Calculated PaDEL descriptors: 1444 (4 sec)
Predicted structures: 10 (0 sec)
Total processing time: 45.69 seconds.

OPERA
OPEn (q)saR App

Please Wait

Initializing...

Please Wait

Calculating descriptors: PaDEL 2D

Please Wait

Checking PaDEL descriptors

Success!

Calculations done in: 45.69 seconds.

OK

OPERA – OPEn (q)saR App: Output

MoleculeID
LogWS_exp
LogWS_pred
AD_WS
AD_index_WS
Conf_index_WS
LogWS_CAS_neighbor_1
LogWS_CAS_neighbor_2
LogWS_CAS_neighbor_3
LogWS_CAS_neighbor_4
LogWS_CAS_neighbor_5
LogWS_InChiKey_neighbor_1
LogWS_InChiKey_neighbor_2
LogWS_InChiKey_neighbor_3
LogWS_InChiKey_neighbor_4
LogWS_InChiKey_neighbor_5
LogWS_DTXXSID_neighbor_1
LogWS_DTXXSID_neighbor_2
LogWS_DTXXSID_neighbor_3
LogWS_DTXXSID_neighbor_4
LogWS_DTXXSID_neighbor_5
LogWS_DSSTOXMPID_neighbor_1
LogWS_DSSTOXMPID_neighbor_2
LogWS_DSSTOXMPID_neighbor_3
LogWS_DSSTOXMPID_neighbor_4
LogWS_DSSTOXMPID_neighbor_5
LogWS_Exp_neighbor_1
LogWS_Exp_neighbor_2
LogWS_Exp_neighbor_3
LogWS_Exp_neighbor_4
LogWS_Exp_neighbor_5
LogWS_pred_neighbor_1
LogWS_pred_neighbor_2
LogWS_pred_neighbor_3
LogWS_pred_neighbor_4
LogWS_pred_neighbor_5



Example1_OPERA2.3_Pred_LogP.csv



Example1_OPERA2.3_Pred_WS.csv



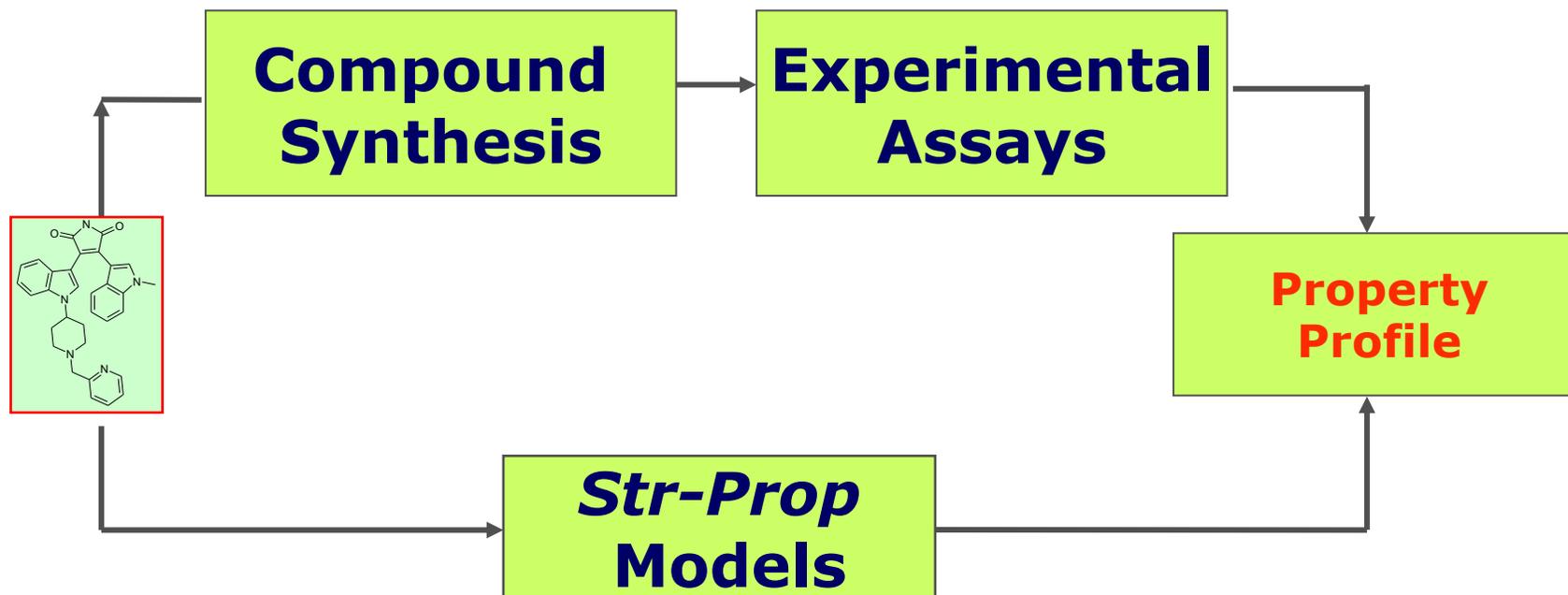
Example1_OPERA2.3_Pred_StrP.csv

MoleculeID	LogWS		AppDom (AD)	AD_index	Conf_index	Neighbor 1		Neighbor 2		Neighbor 3		Neighbor 4		Neighbor 5	
	Exp	Pred				Exp	Pred	Exp	Pred	Exp	Pred	Exp	Pred		
DTXSID1025853	1.119	1.090	1	0.856	0.841	1.119	0.920	0.983	0.641	1.119	1.049	0.527	0.587	0.935	0.508
DTXSID7021100	NaN	2.560	1	0.550	0.550	2.844	2.278	2.862	2.212	4.746	-3.868	-2.826	-3.384	-2.188	-3.800
DTXSID5020449	Structural properties (StrP v2.0) included by default.									11	-1.569	-1.754	-1.490	-2.280	-1.498
DTXSID6020690	MolWeight, nbAtoms, nbHeavyAtoms, nbC, nbO, nbN, nbAromAtom, nbRing,									53	-4.143	-4.505	-4.750	-4.301	-4.743
DTXSID9021976	nbHeteroRing, Sp3Sp2HybRatio, nbRotBd, nbHBdAcc, ndHBdDon,									39	-0.964	-0.476	-0.987	-0.475	-1.712
DTXSID3020415	nbLipinskiFailures, TopoPolSurfAir, MolarRefract, CombDipolPolarizability									01	-2.073	-1.165	-1.461	-2.084	-2.106
DTXSID3039242	-1.640	-1.872	1	0.880	0.725	-1.640	-2.479	-2.790	-2.542	-2.075	-2.264	-2.727	-2.009	-2.625	-2.518
DTXSID0020819	-5.647	-5.541	1	0.864	0.712	-5.647	-4.972	-5.037	-5.223	-5.239	-5.494	-5.428	-2.864	-4.370	-4.760
DTXSID2021238	0.814	0.489	1	0.905	0.719	0.814	-0.153	-0.185	0.424	0.326	0.006	0.622	-0.687	-0.268	-0.998
DTXSID8022292	NaN	-6.202	1	0.546	0.427	-7.217	-4.776	-6.320	-6.199	-8.161	-4.906	-3.226	-6.167	-6.014	-5.734

More details: QMRF registered in the European Commission's Joint Research Center (JRC) QMRF Inventory

https://www.researchgate.net/publication/316789777_QMRF_-_Title_WS_model_for_water_solubility_prediction_from_OPERA_models?channel=doi&linkId=59e624b50f7e9b4f49a97116&showFulltext=true

Experiments are too hard: Wet vs In Silico



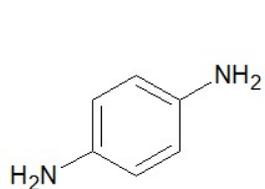
Experience

**"Experience is not what happens to you;
it's what you do with what happens to you."**

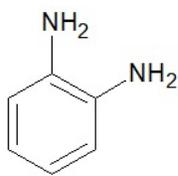
Aldous Huxley (1894-1963)

Models as In Silico Instruments

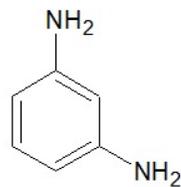
$$P = f(S)$$



Noncarcinogen



Carcinogen



Noncarcinogen

Structure	Name	Boiling point (°C)
	pentane	36.0
	2-methylbutane	27.9
	2,2-dimethylpropane	9.5

P: Property (Physchem, toxicity endpoint, ADME)

f: Mathematical function

S: Structure quantifier(s)

- Inexpensive
- Extremely fast
- Reduce animal use
- Rationalize synthesis and testing
- Help design safer compounds

Structure Quantification: Molecular E-State

$$\delta^v = (Z^v - H) / (Z - Z^v - 1)$$

$$I_i = (\delta^v + 1) / \delta$$

$$\Delta I_i = \sum [(I_i - I_j)] / r_{ij}^2$$

$$E_i = I_i + \Delta I_i$$

LB Kier and LH Hall, *Molecular Structure Description –The Electrotopological State*, Academic Press, 1999.

Models as iHTS: Minimizing Assays (Permeability)

Caco-2 (% Transport Distributions)				
Transport (%)	Category	ESMol < 55	ESMol < 50	ESMol < 45
< 4	Low	324	153	68
> 4	Not Low	2989	2002	1226
Total		3313	2155	1294
% Not Low		90.2%	92.9%	94.7%

Models as iHTS: Minimizing Assays (Solubility)

Kinetic Solubility Distributions

Solubility	Category	ESMol < 50	ESMol < 45	ESMol < 40
< 120	Low	541	235	65
> 120	High	3641	2343	1197
Total		4182	2578	1262
% High Sol		87.1%	90.9%	94.8%

Models as iHTS: Well-known “Rules”

Lipinski's rule of 5 for orally active drugs:

- No more than 5 hydrogen bond donors (the total number of nitrogen–hydrogen and oxygen–hydrogen bonds)
- No more than 10 hydrogen bond acceptors (all nitrogen or oxygen atoms)
- A molecular mass less than 500 daltons
- An octanol-water partition coefficient ($\log P$) that does not exceed 5

Veber's Rule for orally active compounds:

- 10 or fewer rotatable bonds and
- Polar surface area no greater than 140 Å²

Rule of three (RO3) for defining lead-like compounds:

- Octanol-water partition coefficient $\log P$ not greater than 3
- Molecular mass less than 300 daltons
- Not more than 3 hydrogen bond donors
- Not more than 3 hydrogen bond acceptors
- Not more than 3 rotatable bonds

Ghose Filter for druglikeness:

- Partition coefficient $\log P$ in -0.4 to $+5.6$ range
- Molar refractivity from 40 to 130
- Molecular weight from 180 to 480
- Number of atoms from 20 to 70 (includes H-bond donors and H-bond acceptors)

• As with many other rules of thumb, there are many *exceptions*.

Detailed Structure Quantification: Packages

Yap, C. W. (2011) **PaDEL-descriptor**: An open-source software to calculate molecular descriptors and fingerprints. J. Comput. Chem. 32, 1466.

Hong, H., Slavov, S., Ge, W., Qian, F., Su, Z., Fang, H., ... Tong, W. (2012). **Mold2Molecular Descriptors** for QSAR. Statistical Modelling of Molecular Descriptors in QSAR/QSPR, 65–109. doi:10.1002/9783527645121.ch3

Moriwaki, H., Tian, Y.-S., Kawashita, N., and Takagi, T. (2018) **Mordred: a molecular descriptor calculator**. J. Cheminf. 10, 4.

... and there are several commercial packages

Predictive Model: A Computational Instrument

$$\mathbf{P} = \mathbf{f} (\mathbf{S})$$

P: Property (Physchem, tox., ADME)

f: Mathematical function

S: Structure quantifier(s)

Black boxes vs Interpretable Models

$$\log P = -1.167 \times 10^{-4}S^2 - 6.106 \times 10^{-2}S + 14.87Ov^2 - 43.67Ov + 0.9986I_{\text{alkane}} + 9.57 \times 10^{-3}Mw - 0.13D - 4.929Q_{ON} - 12.17Q_N^4 + 26.81Q_N^2 - 7.416Q_N - 4.551Q_O^4 + 17.92Q_O^2 - 4.03Q_O + 27.273$$

$$n = 118, r = 0.9388, F = 115.1$$

S: Molecular surface

Ov: Ovality of the molecule*

I_{alkane}: Indicator variable for alkanes

Mw: Molecular weight

D: Calculated dipole moment

Q_{ON}: Sum of absolute values of atomic charges on nitrogen and oxygen atoms

Q_N: Square root of the sum of the squared charges on nitrogen atoms

Q_O: Square root of the sum of the squared charges on oxygen atoms (AM1)

**J. Am. Chem. Soc.* 1989, 111, 3783.

Models with Structural Features as Descriptors

Can point out to a toxicologist:

a substructure potentially rendering a molecule toxic/nontoxic

May indicate to a pharmacologist:

the structural moiety required for certain pharmacological action

May help a biochemist:

generate hypothesis for a possible mode of action

Can guide a chemist:

in designing “better” compounds by combining substructures

Saagar - A new, extensible set of molecular substructures for interpretable QSARs and read-across applications

Sedykh AY; Shah RR; Kleinstreuer NC; Auerbach SS; Gombar VK (2021). "Saagar-A New, Extensible Set of Molecular Substructures for QSAR/QSPR and Read-Across Predictions." *Chemical Research in Toxicology* 34(2):634-640

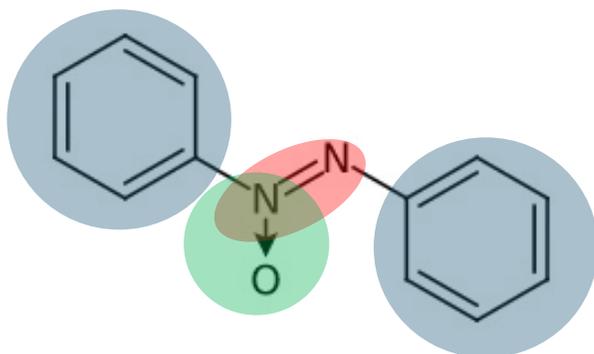
Saagar Features as Structure Descriptors

- A parsimonious but extensible set of chemistry-aware and chemically viable functional groups and moieties - *Saagar*
- Systematically gathered by studying and reviewing available open-source literature that highlights relationships between substructural moieties and a variety of physico-chemical, ADME, and toxicological properties
- The *Saagar* set encodes and enumerates salient molecular features like:
 - Hierarchical macro structural class (inorganic, aliphatic - acyclic or alicyclic, aromatic, fused or unfused carbo-aromatic and hetero-aromatic, and organometallic),
 - Elemental makeup,
 - Ring size, ring substituents and their positions,
 - Separation of certain hetero atoms to account for key interactions
 - Typical scaffolds for endogenous biochemicals (e.g., amino acids, lipids)
 - Scaffolds present in common medicinal and industrial chemicals
- *Saagar* version SGR-v0120 has 834 features coded as SMARTS strings
- Provide sufficient coverage of chemicals in Drugbank, ChEMBL, and Tox 21 sets

Example: Saagar v1 Collection

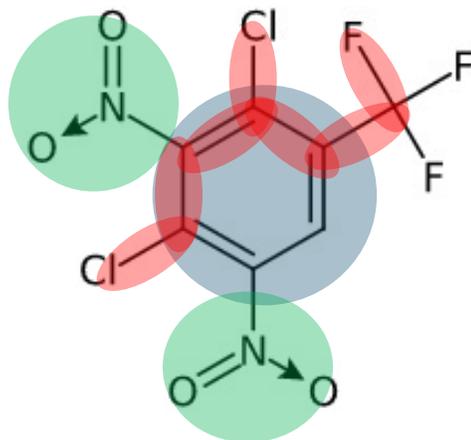
834 Saagar features are coded them as SMARTS* (**S**MILES **A**Rbitrary **T**arget **S**pecification)

DSSTox_CID4555



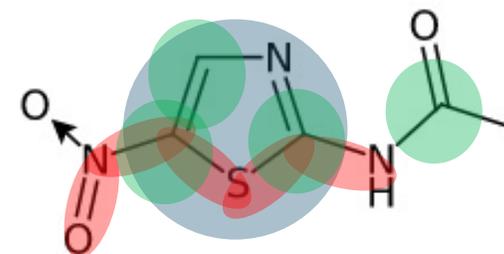
[R] **12**
[N!r]=[N!r]
[N+!\$(N=O)][O-X1]

DSSTox_CID24639



6
[F,Cl,Br,I]~*~*~*~[F,Cl,Br,I]
AR2NO2_META

DSSTox_CID26391



5
[O]~*~*~*~*~[NX3;NH2,\$([NH][#6]),
,\$([N]([#6])[#6]))
[#6;X3]

How good is the *Saagar* Feature Set?

Experiment 1:

How effective *Saagar* features can be for read-across applications?

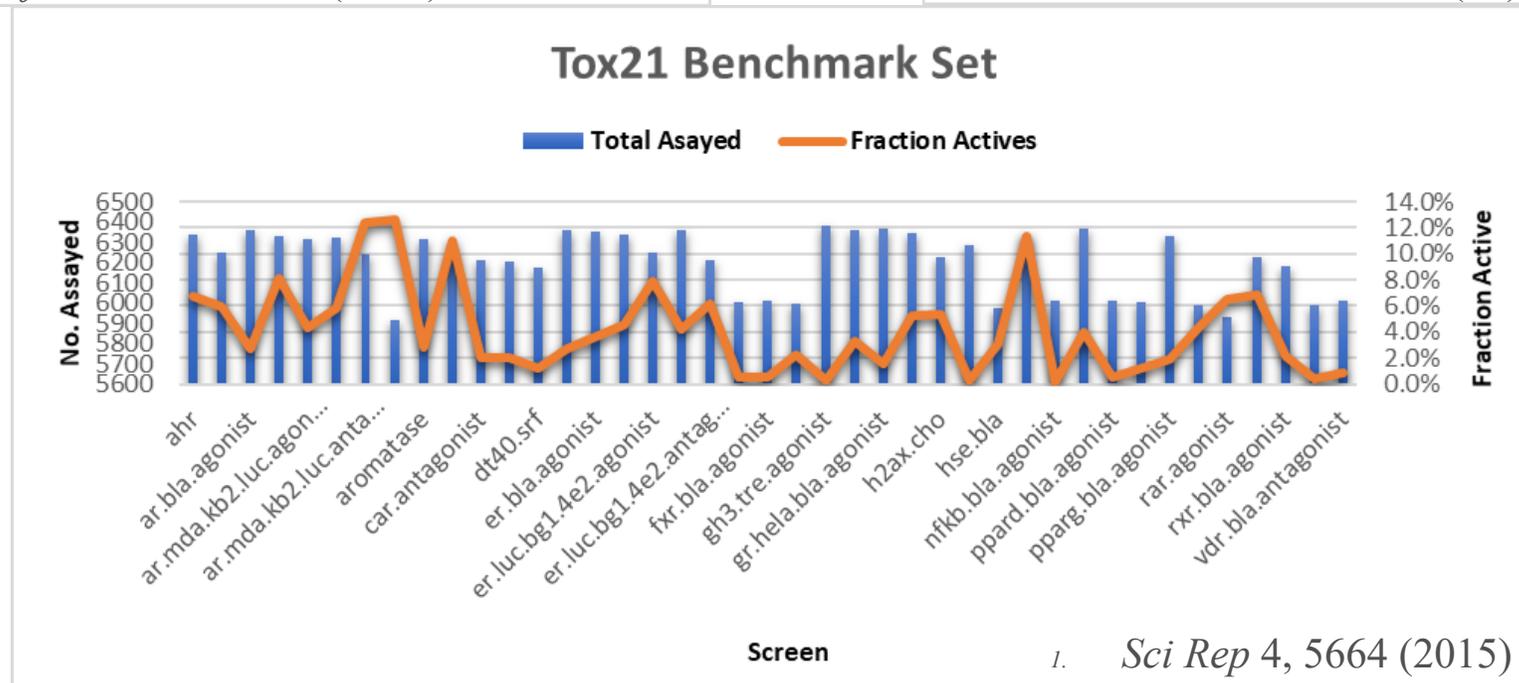
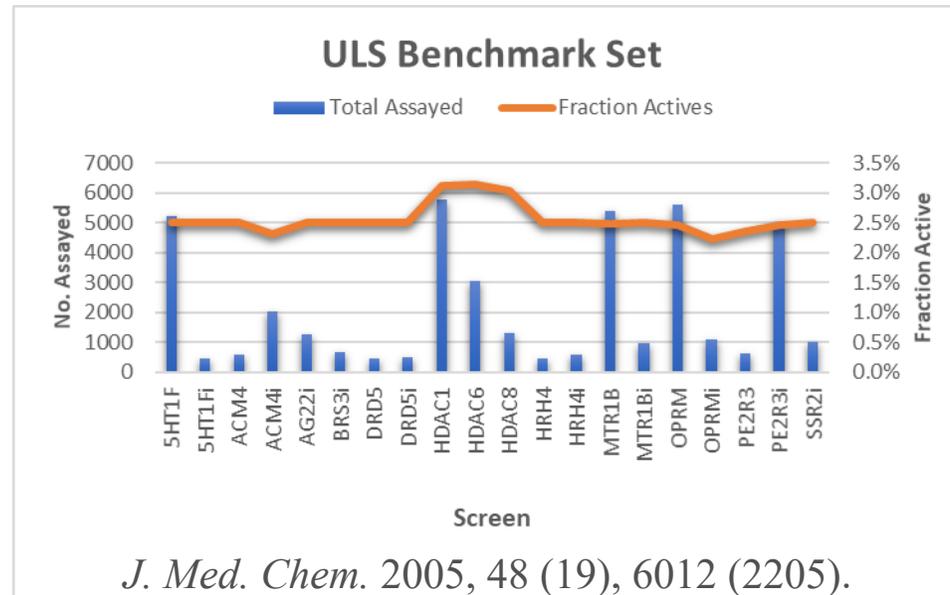
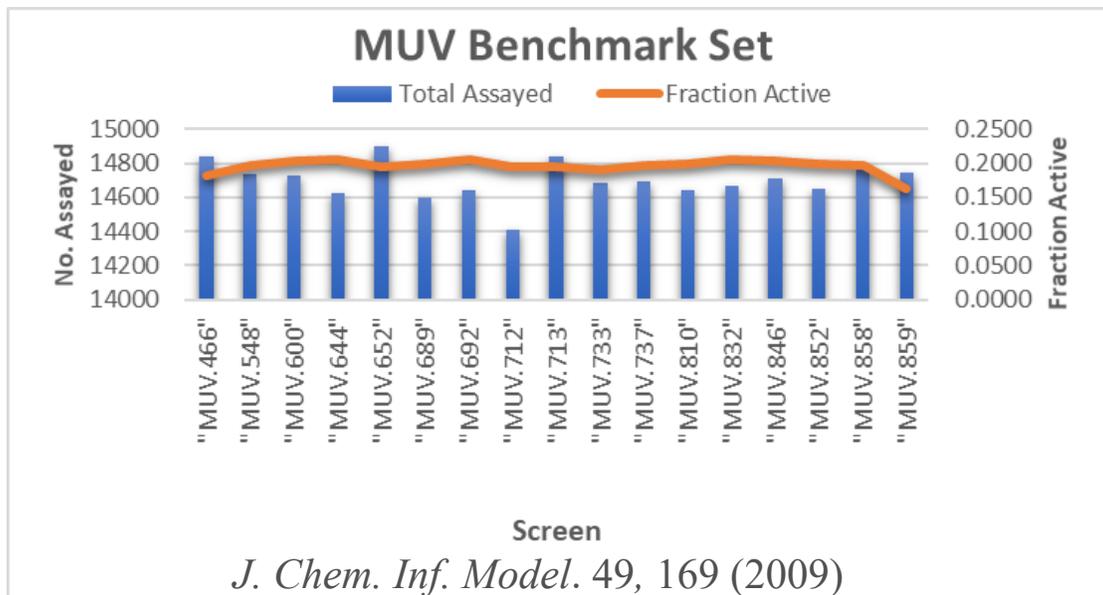
By comparing their active extraction efficiency with other fingerprints

Experiment 2:

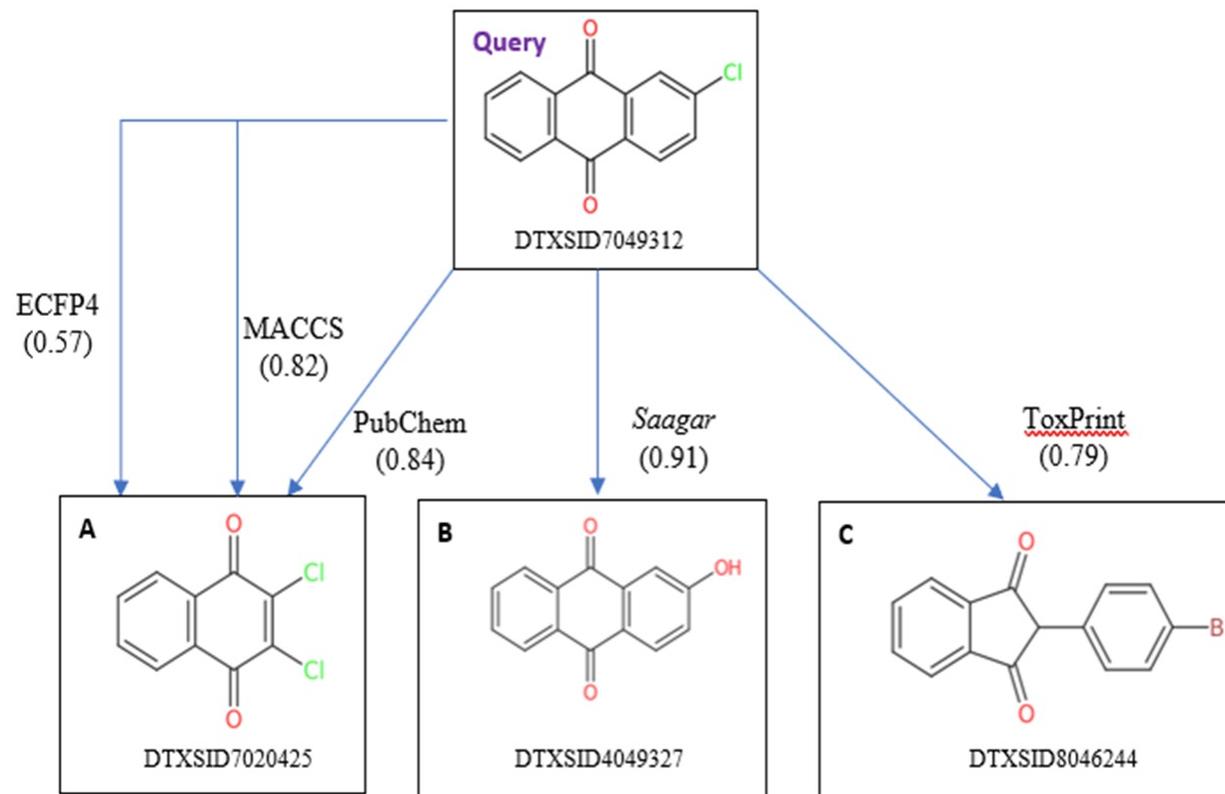
How good and interpretable QSAR models *Saagar* features will yield?

By comparing performance of QSAR models developed with *Saagar* features and with other descriptors.

Analogue Extraction with *Saagar* Features



Chemistry-aware Analogues and Similarity with *Saagar*

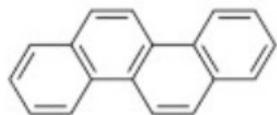


*For more information see www.sciome.com/saagar

Sedykh AY; Shah RR; Kleinstreuer NC; Auerbach SS; Gombar VK (2021). "Saagar-A New, Extensible Set of Molecular Substructures for QSAR/QSPR and Read-Across Predictions." *Chemical Research in Toxicology* 34(2):634-640

Structure Resolution with Saagar

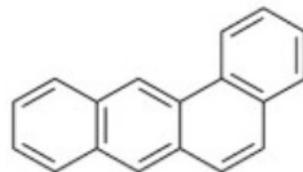
Query



Chrysene
DTXSID:DTXSID0022432
CASRN:218-01-9

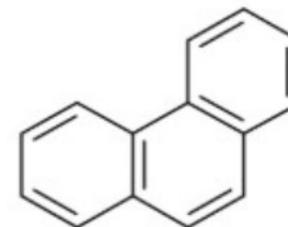
	A	B	C	D	E
ToxPrint	0.50	0.75	0.50	0.60	0.43
ECFP4	0.60	0.86	0.67	0.39	0.67
Pubchem	1.00	0.94	0.97	1.00	0.93
MACCS	1.00	1.00	1.00	1.00	0.88
Saagar	0.98	0.76	0.81	0.88	0.72

A



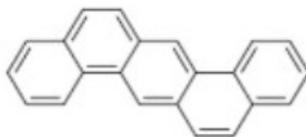
Benz(a)anthracene
DTXSID:DTXSID5023902
CASRN:56-55-3

B



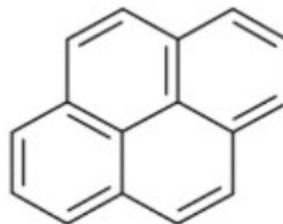
Phenanthrene
DTXSID:DTXSID6024254
CASRN:85-01-8

C



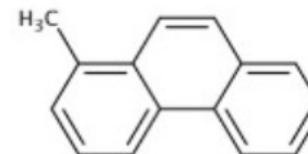
Dibenz(a,h)anthracene
DTXSID:DTXSID9020409
CASRN:53-70-3

D



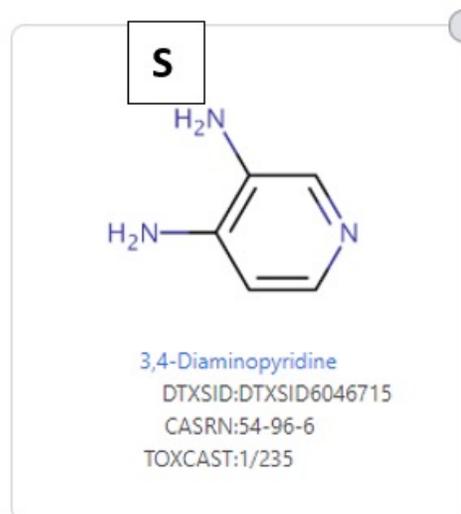
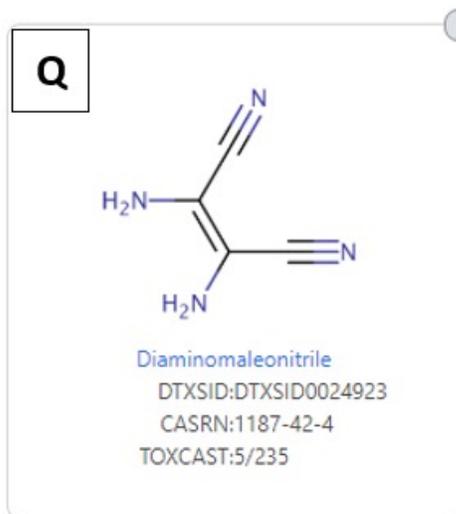
Pyrene
DTXSID:DTXSID3024289
CASRN:129-00-0

E



1-Methyl phenanthrene
DTXSID:DTXSID6025648
CASRN:832-69-9

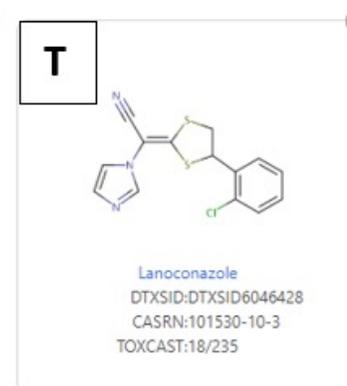
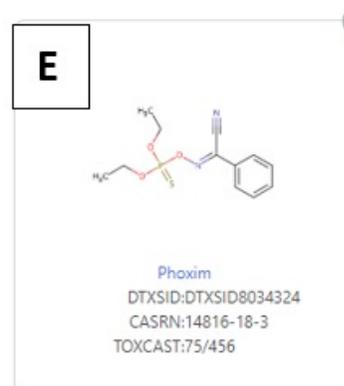
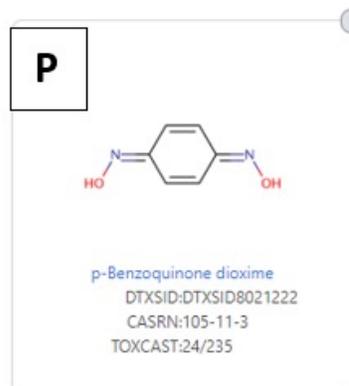
Extensibility of Saagar



Fingerprint	Window	Similarity to Query
<u>ToxPrint</u>	T	0.29
ECFP	E	0.12
PubChem	P	0.31
MACCS	M	0.41
Saagar	S	0.33

[N;H2][C;!R]=[C;!R] and/or

N#C[C;!R]=[C;!R].



***Saagar* Performance vs Mordred in Predictive Models**

- Developed predictive models for outcomes in 41 Tox21 Assays using *Saagar* and Mordred
- In 5-fold cross validation test, the average absolute difference in AUROC between *Saagar* and Mordred models was just 0.02
- In a detailed analysis comparing accuracy, NPV, PPV, and AUROC for five representative sets of the 41 Tox21 data sets,
 - *Saagar* performed better than Mordred in 14 of the 20 comparisons,
 - *Saagar* performed as good as Mordred in 3 of the 20 comparisons, and
 - *Saagar* performed a bit inferior to Mordred in 3 of the 20 comparisons.

Saagar descriptors yield models better than or as good as the models using Mordred descriptors
– **and chemistry-backed reasoning for prediction**

Methods to Develop Predictive Models

$$P = f(S)$$

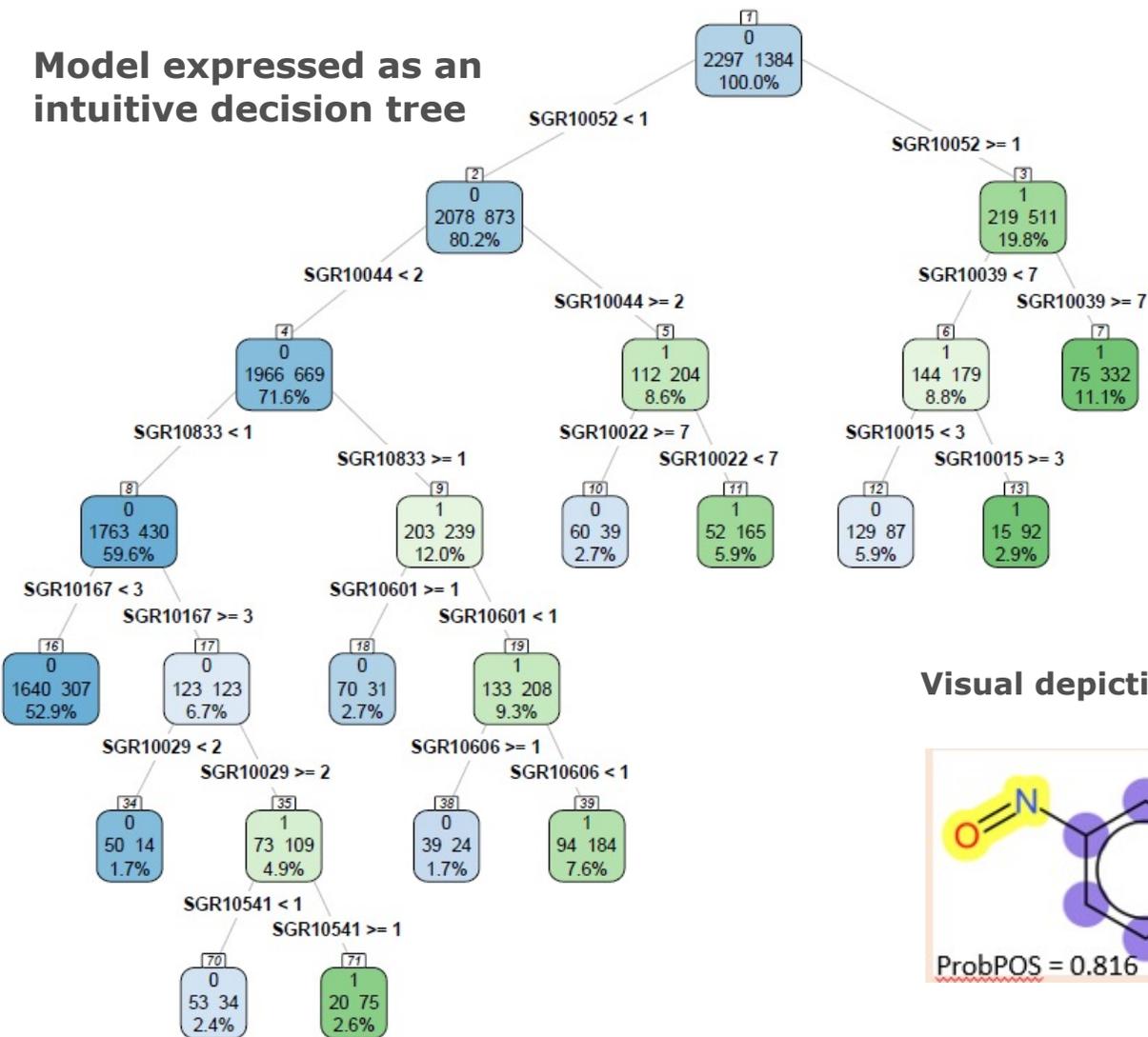
Modeling Methods:

- Regression Analysis
- Linear and Canonical Discriminant Analysis
- Partial Least Squares
- Principal Component Regression
- Nearest Neighbor Analysis
- Neural Networks
- Inductive Logic
- Support Vector Machines
- **Recursive Partitioning** and Random Forest

QSAR Methods, Giuseppina Gini. In *In Silico Methods for Predicting Drug Toxicity*
Volume 1425 of the series *Methods in Molecular Biology* pp 1-20, 17 June 2016

Saagar-RP Model: Ames Test (-S9)

Model expressed as an intuitive decision tree



Model expressed as structure-based rules

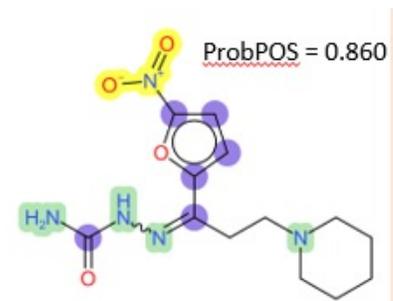
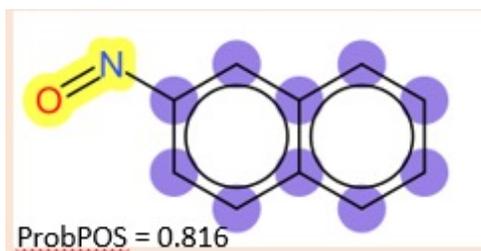
Prob_POS	Rule Hits	R			
0.816	407	SGR10052 >= 1	&	SGR10039 >= 7	
0.860	107	SGR10052 >= 1	&	SGR10039 < 7	& SGR10015 >= 3
0.403	216	SGR10052 >= 1	&	SGR10039 < 7	& SGR10015 < 3
0.760	217	SGR10052 < 1	&	SGR10044 >= 2	& SGR10022 < 7
0.394	99	SGR10052 < 1	&	SGR10044 >= 2	& SGR10022 >= 7

7 more rows

Saagar Code	Hit Freq (3681)	Positive/Negative	Feature Description
SGR10015	2320	972/1348	[#7]
SGR10022	2837	1024/1813	[CX4]
SGR10029	3599	1376/2223	[!#6]
SGR10039	3266	1249/2017	[#6;X3]
SGR10044	325	211/114	[r3,r4]
SGR10052	731	512/219	[OH0]~N

5 more rows

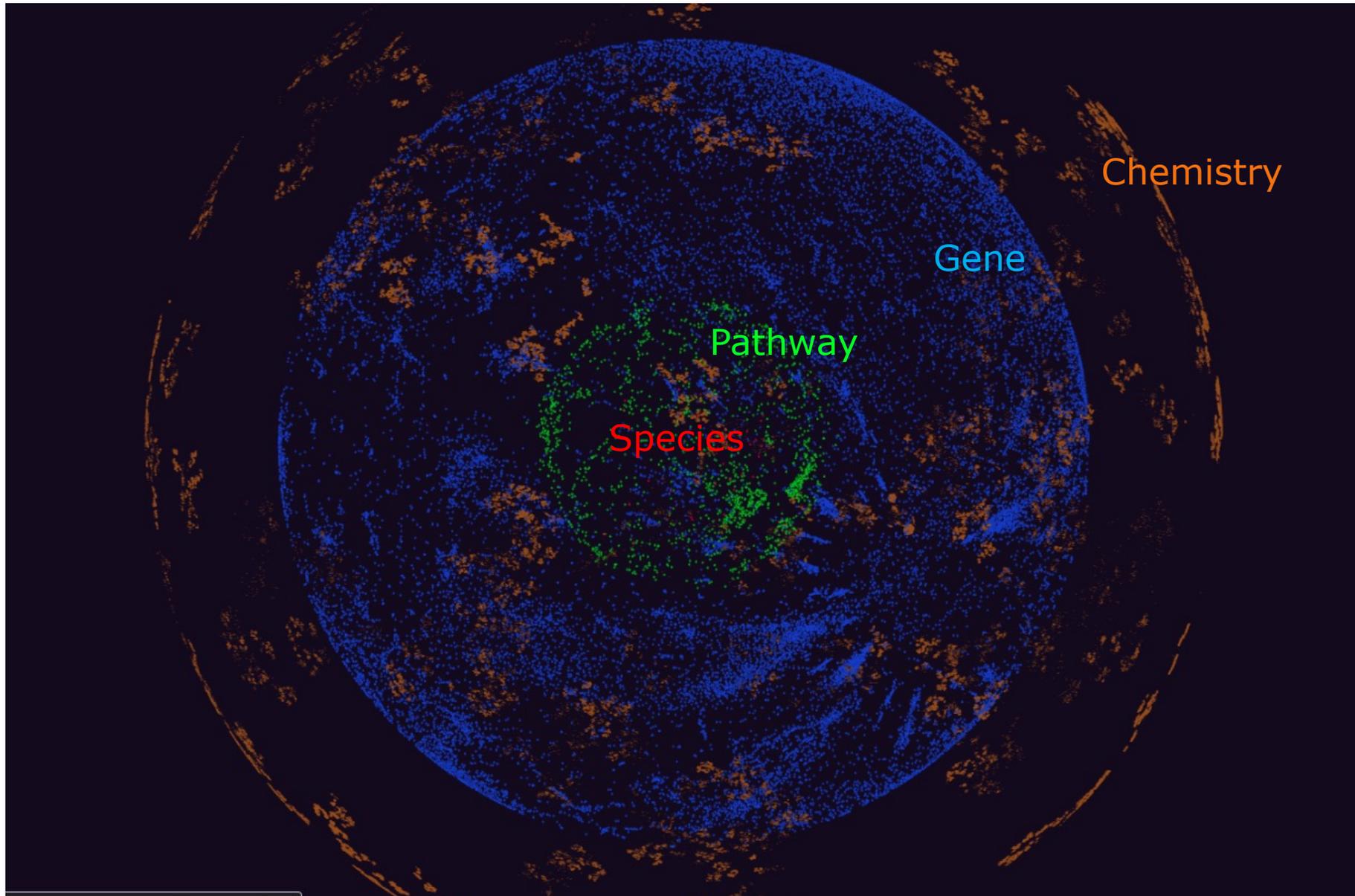
Visual depiction for chemistry-backed reasoning





OrbiTox - An integrated framework for concerted view and connectivity among multi-domain data, predictive models, and cheminformatics tools

OrbiTox Big Data Organization





Chemistry Orbit:

- ~ **900,000 chemical substances** (name, structure, DSSTOX ID, macro class, etc.)
- ~ **1,400 chemicals** with carcinogenic potency data (CPDB)
- ~ **400 chemicals** with human carcinogenicity group (IARC)
- ~ **600 chemicals** with sex/species-specific carcinogenicity evidence (NTP Technical Reports)
- ~ **4,000 chemicals** with bacterial mutagenicity calls (ToxNet)

Gene Orbit

~ **25,000 annotated human genes:**

GO Terms, Synonyms, Name, Chromosome, Location, etc.

Pathway Orbit

~ **2,000 annotated pathways**

Name, Size, Set, DB

~40,000 connections with genes

Species Orbit

~ **200 organisms with life cycle information**

Genus, family, class, genome, life span, weight, etc.

~1,800 Connections with carcinogenicity

OrbiTox Data Connectivity (Chemical Query)



DTXSID2020139

Data List Recent

DTXSID2020139

Data Structure Connections

DTXSID2020139 1/1 < > ⌵ ×

Name	Benzo(a)pyrene
CAS	50-32-8
PubChem CID	2336
SMILES	<chem>c1ccc2c(c1)cc1ccc3ccccc4ccc2c1c34</chem>
Class	8
Class Name	Polyaromatic
MW	252.0939004
Heavy Atoms	20
Bonds	24

DTXSID2020139

Data Structure Connections

AHR
aryl hydrocarbon receptor

RARA
retinoic acid receptor alpha

Human
Carcinogen

Mouse
Carcinogen

Rat
Carcinogen

ID	AHR
Layer	GENE
Size	0
Uniprot ID	P35869
Entrez Gene ID	196
REF	https://pubchem.ncbi.nlm.nih.gov/bios
LINK	AHR
ID	DTXSID2020139
INFO	Tox21 Program, hAHR agonist
TYPE	Activator

ID	Rat
Layer	LIFEFORM
Size	8.7
Class	Mammalia
Life Cycle	2-3 yrs
Length	20-50 cm
REF	CPDB
LINK	Rat
ID	DTXSID2020139
INFO	(TD50 = 0.956)
TYPE	Carcinogen

Filter

OrbiTox Data Connectivity (Pathway Query)



FRUCTOSE_METABOLISM

Data Connections

FRUCTOSE_METABOLISM 1/1 < > v x

Layer	PATHWAY
Size	7
Full Name	REACTOME_FRUCTOSE_METABOLISM
DB	REACTOME
SRC	http://www.gsea-msigdb.org/gsea/msigdb/cards/REACTOME_FRUCTOSE_ME
Set	231 229 3795 6652 26007 216 132158

Data **Connections**

FRUCTOSE_METABOLISM 1/1 < > v x

- AKR1B1
aldo-keto reductase family 1 member B
- ALDH1A1
aldehyde dehydrogenase 1 family member A1
- ALDOB
aldolase, fructose-bisphosphate B
- GLYCK
glycerate kinase
- KHK
ketoheokinase

Data **Connections**

AKR1B1 1/1 < > v x

- DTXSID80163642
Ranirestat
- DTXSID40179264
Lidorestat
- DTXSID80904890
Tolrestat
- DTXSID4031980
Resveratrol
- DTXSID4046654
Fidarestat

Data **Structure**

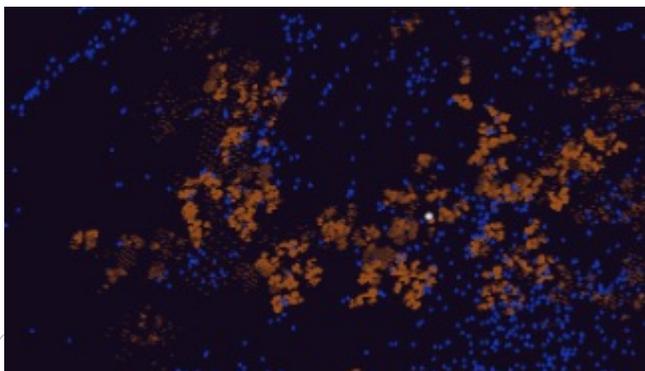
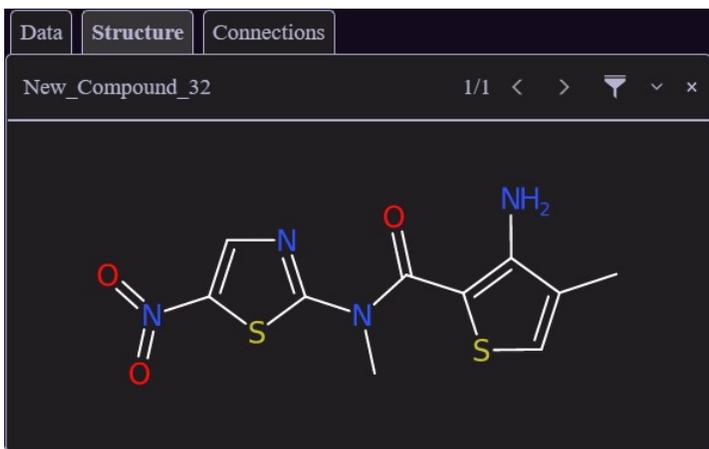
ID	DTXSID80163642
Name	Ranirestat
Layer	CHEM
Size	26
CAS	147254-64-6
REF	NA
LINK	AKR1B1
ID	DTXSID80163642
INFO	IC50 = 15nM
TYPE	Inhibitor

OrbiTox Toxicity Prediction Models



Search: O=N(=O)C1=CN=C(N(C)C(=O)C2=C(N)C(C)=CS2)C

New_Compound_32
MLS000757084
DTXSID20315461
Similarity: 0.78



Data Structure Connections

New_Compound_32 1/1 < > ⌵ ⌵ ×

AHR
Tox21, AHR, 100uM

AR
Tox21, AR Mda-Kb2-Luc-Antagonist (Lower agonist), 100uM

ATAD5
Tox21, ELG1 Luc-Agonist, 100uM

CYP19A1
Tox21, AROMATASE, 100uM

ESR1
Tox21, ER Luc-Bg14e2-Antagonist (Lower agonist), 100uM

Cell Line	HepG2
Model Name	ahr
Reference	https://tripod.nih.gov/tox21/assays/
Probability	0.72
Activity	inactive



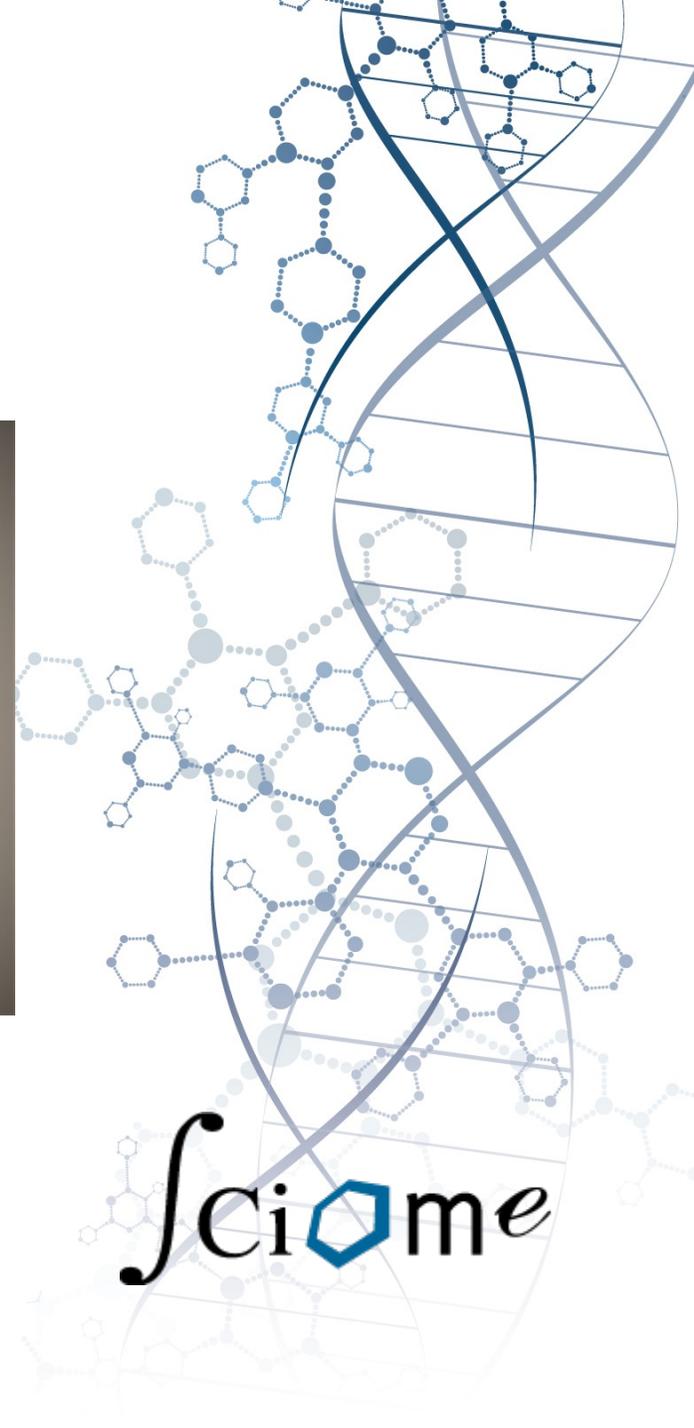


OrbiTox In Action

Austin Ross

Software Developer

Austin.Ross@sciome.com



How to access OrbiTox

<https://apps.sciome.com/orbitox/>

Username:

guest

Password:

guest

Why Develop Predictive Models?

Toxicity Tests for Regulatory Acceptance (OECD)

- Acute Oral, inhalation, and dermal toxicity
- Genotoxicity (*in vitro*, *in vivo*)
- Reproductive Toxicity
- Developmental Toxicity
- Organ Toxicity (hepato-, cardio-, and nephrotoxicity)
- Skin and eye irritation and skin sensitization
- Carcinogenicity
- Bioaccumulation and biodegradation
- Acute aquatic toxicity (fish, daphnia, algae)

Safety Tests for Chemicals Entering Commerce:

- Industrial chemicals
- Pesticides and Insecticides
- Cosmetics
- Drugs
- Food additives, etc.

TSCA list alone contains 67,385 chemicals

HPV (> 1mil lbs/yr) list has over 2200 chemicals

EU's COSMOS project has 5500 cosmetic ingredients

Why Develop Predictive Models?

Regulatory bodies are requiring it

- The ICH M7 guidelines state: “Two (Q)SAR prediction methodologies that complement each other should be applied. One methodology should be expert rule-based and the second methodology should be statistical-based.”
- One of the aims of REACH (**R**egistration, **E**valuation, **A**uthorisation and Restriction of **C**hemicals) regulations is to: “...Promote alternative methods for the assessment of hazards of substances ”
- Fund for the Replacement of Animals in Medical Experiments (FRAME) ...to assess the prospects of developing *in vitro*, target organ, and theoretical model systems which would lead to reduction in studies on live animals (3Rs):
- Reduction of the number of animals used, Refinement of the endpoints of animal experiments, Replacement of animals by other techniques

Experimental Assessment is Costly

Estimated External Costs for Representative Nonclinical Safety Studies

<i>Study type</i>	<i>Approximate external costs^a (U.S. \$)</i>			<i>Approximate duration^b (wk)</i>
	<i>Rats</i>	<i>Dogs</i>	<i>Monkeys</i>	
Single-dose and range-finding				
Single dose	6000–25,000	20,000–50,000	20,000–70,000	10–12
Combined Single and 7-d repeat dose	20,000–80,000	40,000–90,000	80,000–120,000	10–12
Combined single and 10-d repeat dose	35,000–67,000	45,000–75,000	85,000–125,000	10–12
Repeat dose toxicology				
7-d	20,000–35,000	34,000–70,000	55,000–75,000	14–16
14-d	40,000–115,000	90,000–130,000	100,000–190,000	14–16
28-d	70,000–150,000	80,000–195,000	150,000–300,000	16–18
3-mo	110,000–270,000	165,000–200,000	240,000–500,000	30–34
6-mo	215,000–350,000	190,000–300,000	350,000–500,000	40–44
9-mo	275,000–375,000	250,000–500,000	400,000–620,000	52–56
12-mo	320,000–490,000	320,000–470,000	500,000–840,000	68–74
Genetic toxicity				
Ames test		5000–10,000		8–12
In vitro chromosomal aberration assay		10,000–35,000		12–16
In vivo micronucleus test (mice/rats)		10,000–30,000		12–16

Anticancer Drug Development Guide: Preclinical Screening, Clinical Trials, and Approval, Editors: **Teicher**, Beverly A., **Andrews**, Paul A. (Eds.)

Predictive Models: For Increased Efficiency and Saving Animal Lives

Cost in \$:

Cost/cmpd of typical HT ADME assays: ~\$250, 20min

Number of compounds assayed/year: 12000

Cost/year: \$3,000,000

If we could eliminate just ~15% assays, i.e., assays for 1800 compounds

Savings: 600 hrs of screening time, ~ \$450,000

Cost in animal sacrifice:

According to the Humane Society of the United States:

In total, an estimated 25 million animals are used annually in research, testing, and education in the United States (over 1 million un-bred, mainly for toxicity evaluation)

Conclusions

- In silico experiments are experiments too.
- In silico models or computational instruments save, time, money, and animal lives.
- Tools like EPA's CompTox Chemicals Dashboard, ICE, OPERA, and **OrbiTox** are great resources for bioassay data and predictive models.
- *Saagar* set is an extensible, endpoint-agnostic set of substructures for cheminformatics applications.
- In an elaborate benchmark study, *Saagar* performed better than commonly used fingerprints in extracting analogues for read-across studies.
- **OrbiTox** is one-of-a-kind interactive, translational discovery platform that gives concerted view of large amounts of multi-domain data, connectivity among data, and predictive models.
- Predictive models in **OrbiTox** provide chemistry-backed reasoning for every prediction.
- **OrbiTox** is extensible to add proprietary data, custom models, and new data domains.



Thank you for joining us !

This concludes the Big Data in Environmental Science and Toxicology series

But on-line data science training continues on . . .



<https://training.tamids.tamu.edu/>