The influence of data curation on QSAR Modeling – examining issues of quality versus quantity of data

Kamel Mansouri, Chris Grulke, Ann Richard, and Antony J. Williams*

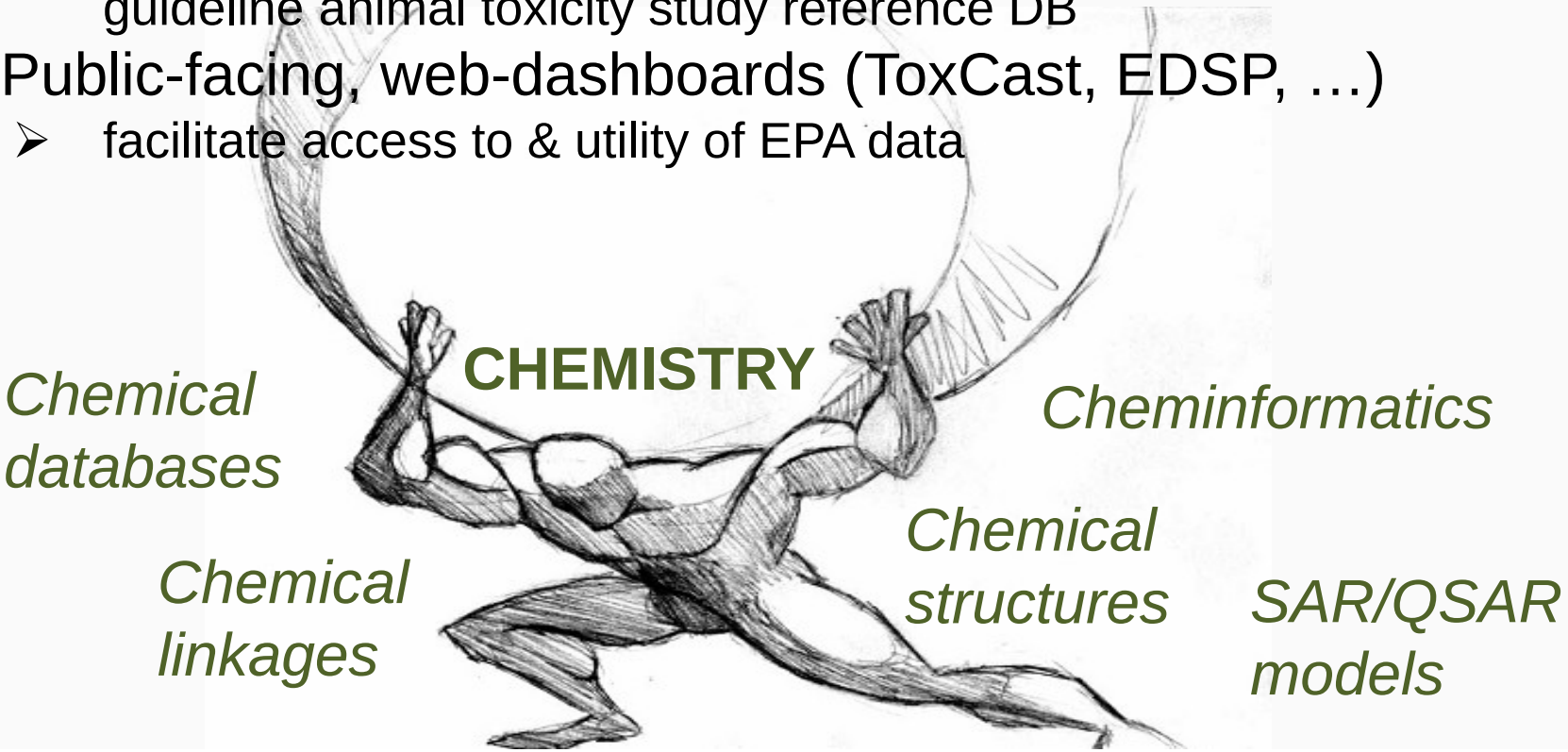
National Center for Computational Toxicology, US EPA, RTP, NC

March, 2016

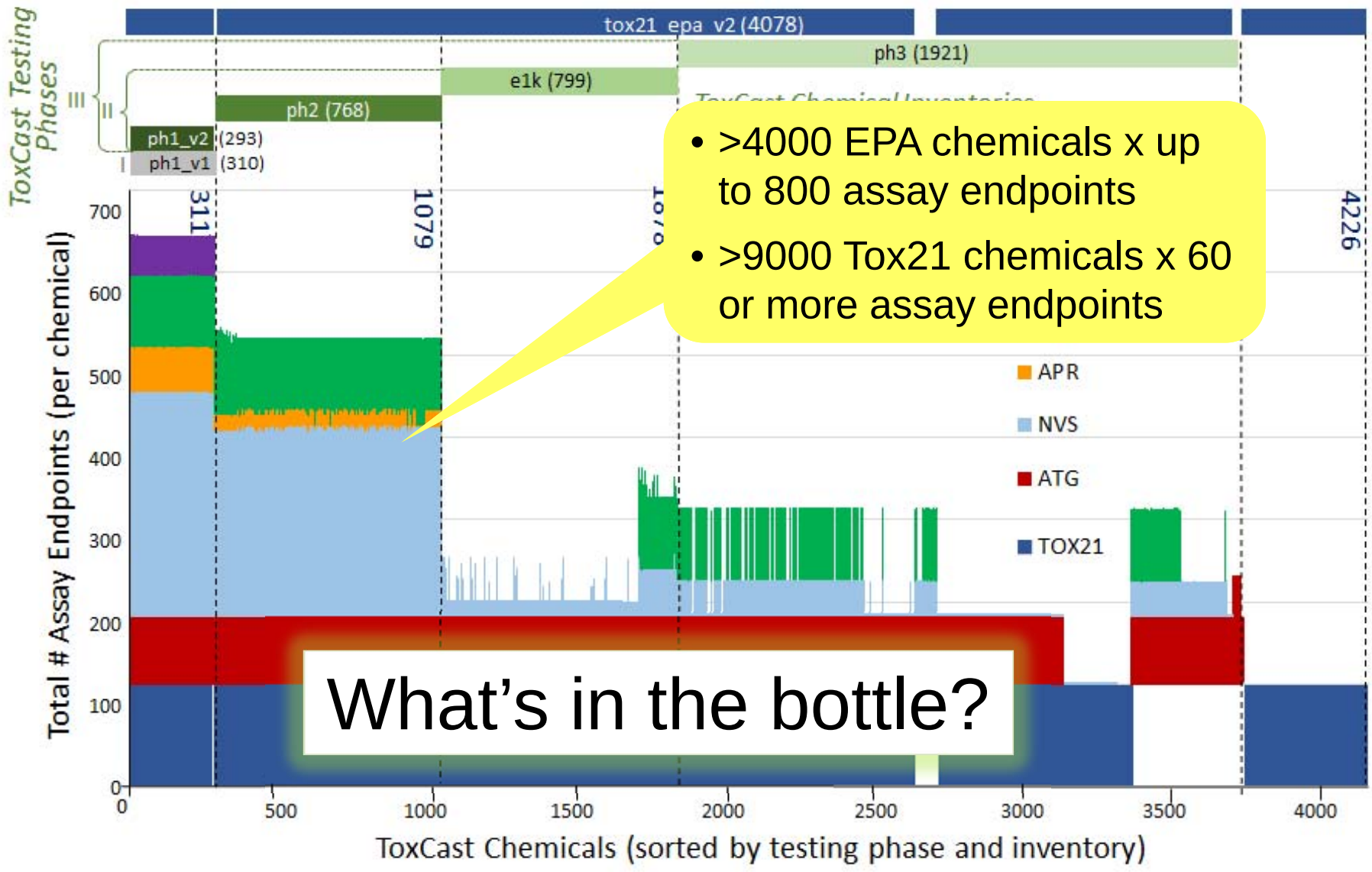
ACS Spring Meeting, San Diego, CA

EPA's National Center for Computational Toxicology (NCCT)

- ToxCast (EPA) & Tox21 (Multi-Agency)
  - screening >3800 (ToxCast) to >10K (Tox21) environmentally relevant chemicals across 10's to 100's of HTS assays
- ACToR, ExpoCast, CPCat, ToxRef DB
  - meshed CAS lists, product-use database, exposure models, guideline animal toxicity study reference DB
- Public-facing, web-dashboards (ToxCast, EDSP, …)
  - facilitate access to & utility of EPA data



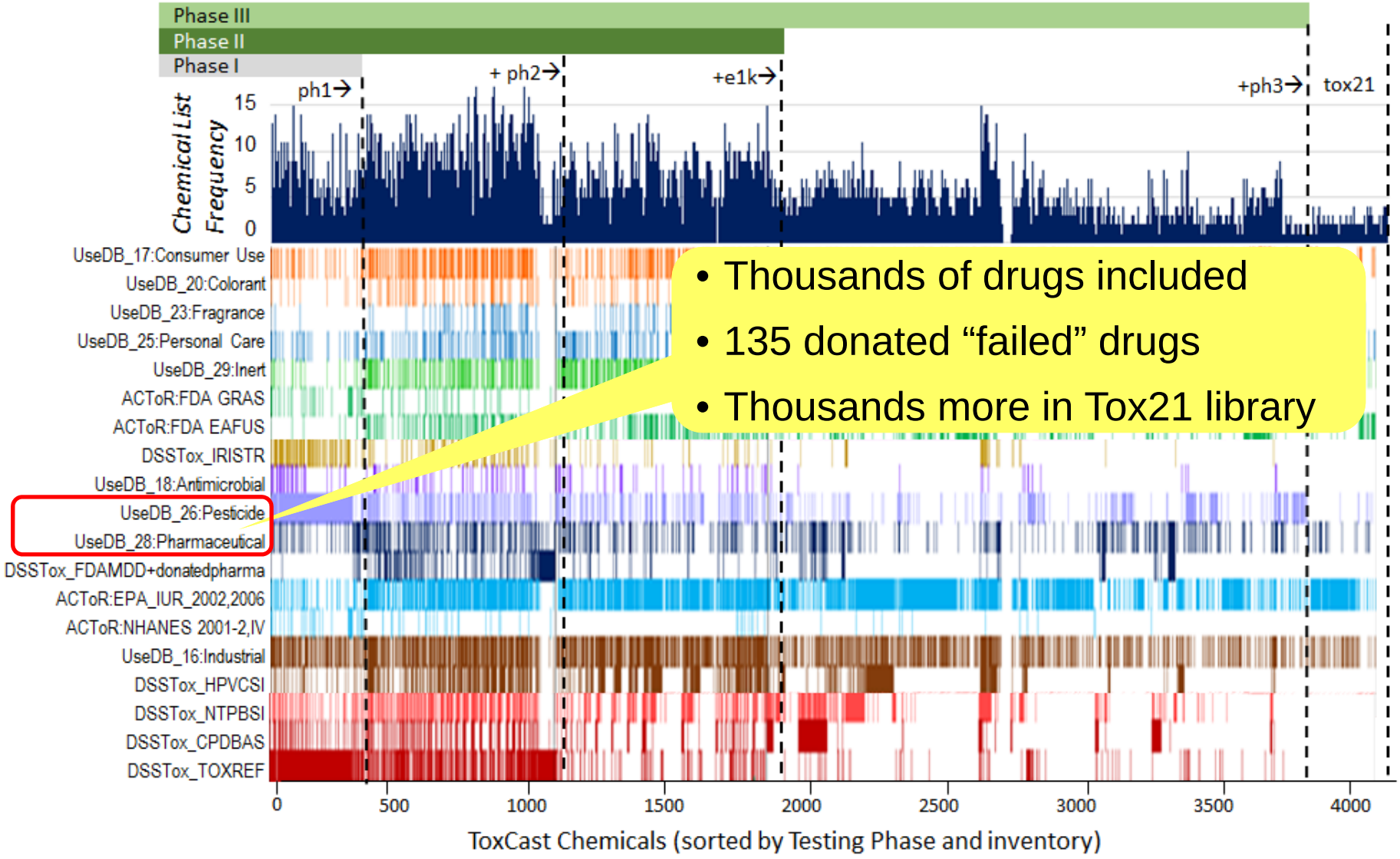
ToxCast chemical x assay counts summary (Jan 2016)



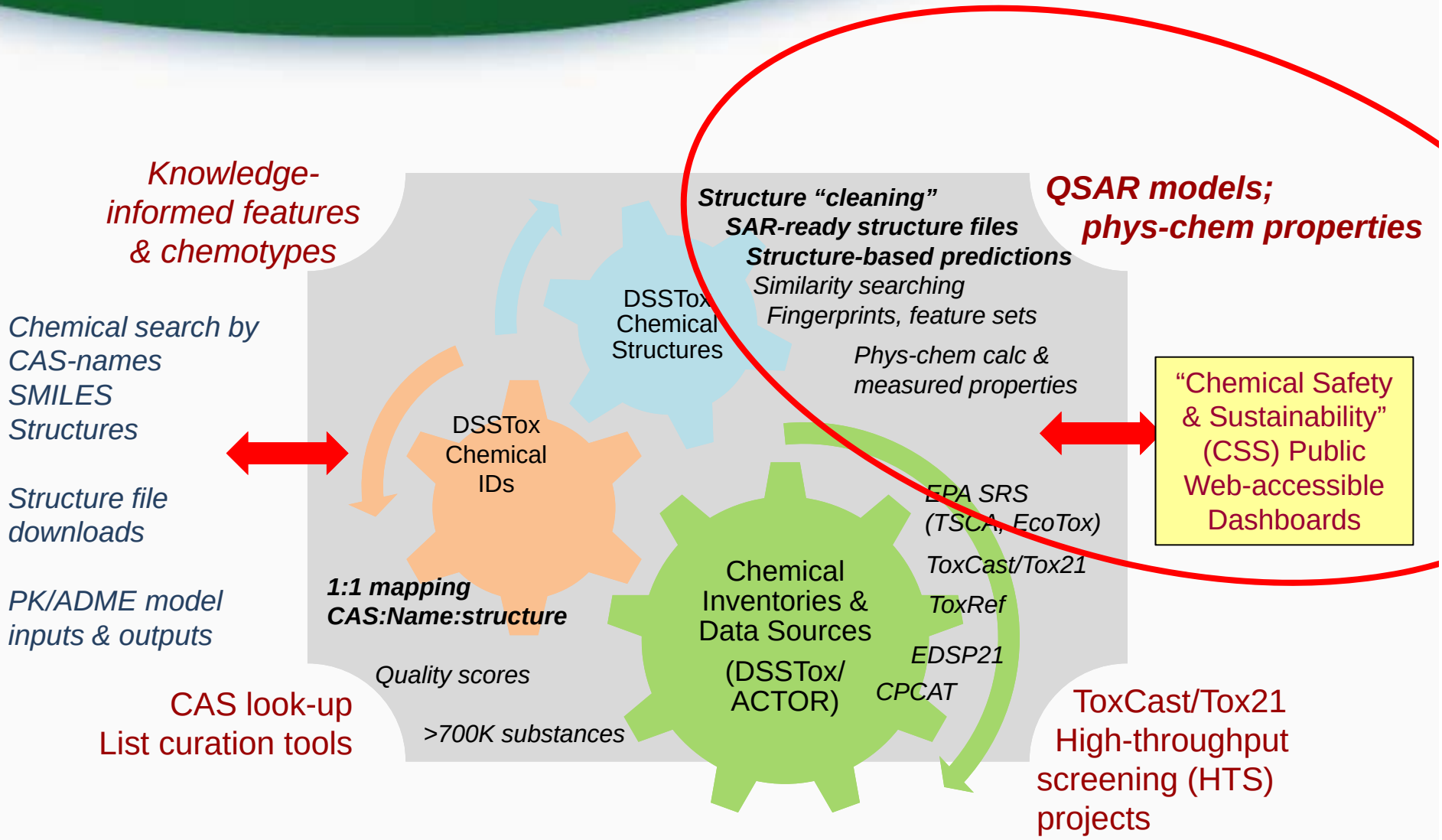
- >4000 EPA chemicals x up to 800 assay endpoints
- >9000 Tox21 chemicals x 60 or more assay endpoints

What's in the bottle?

Tox Data & Use List Coverage of ToxCast Chemicals



Chemistry Foundation to Support Multiple NCCT & EPA Projects



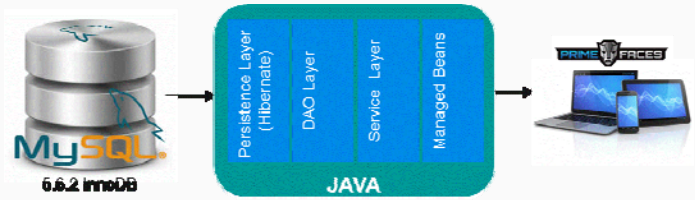
DSSTox Update

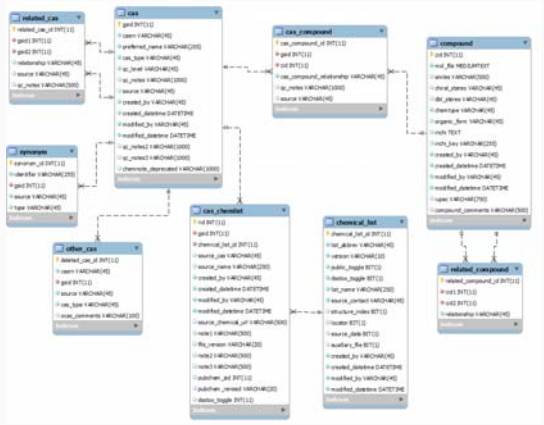
DSSTox_v1

- Manually curated 25K substance records
- EPA-focus, environmental tox datasets
- Emphasis on accurate CAS-name-structure annotations
- Public resource for high-quality structure-data files (SDF)

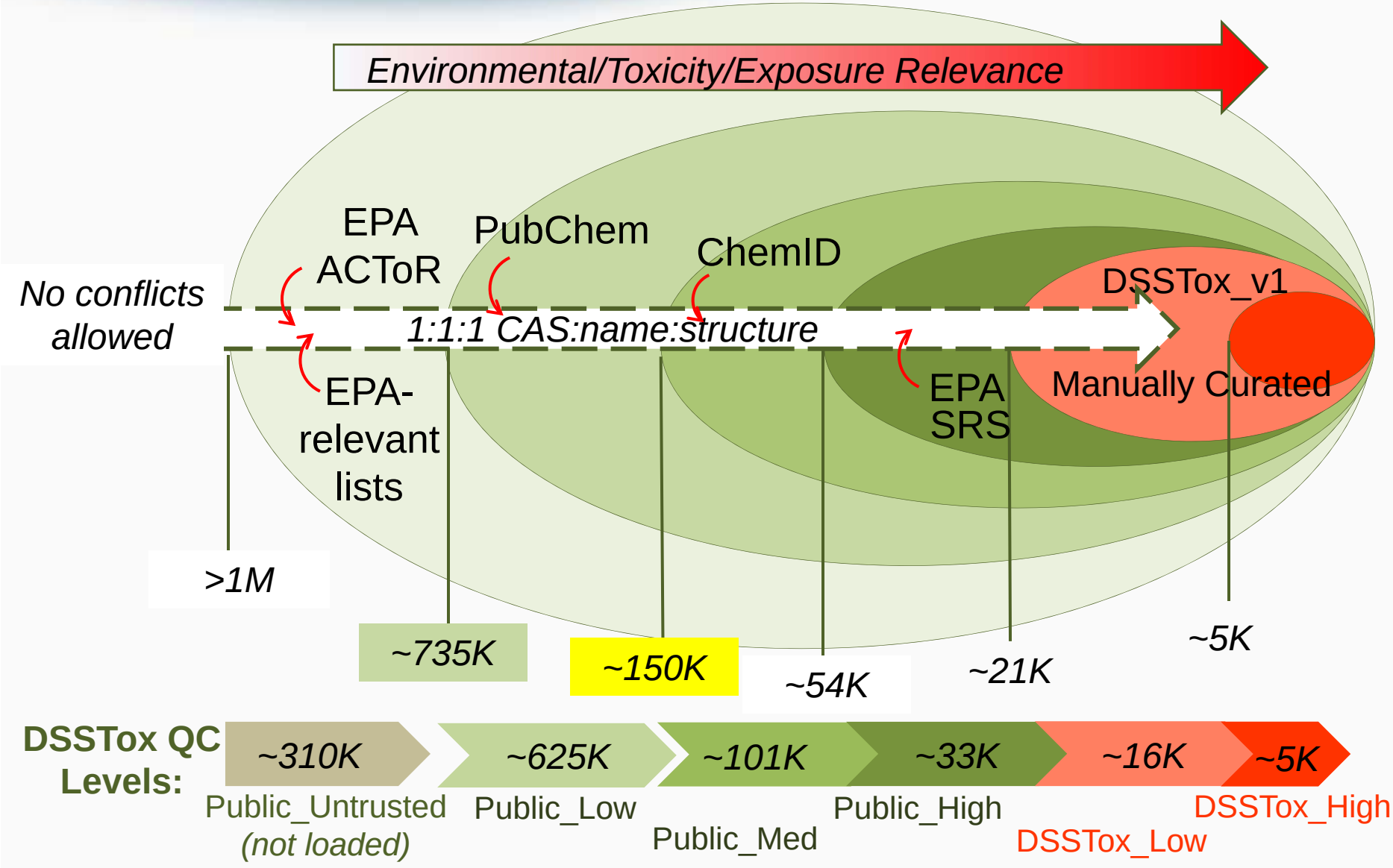
DSSTox_v2

- Convert DSSTox tables to MySQL
- Develop curation interface & cheminformatics workflow
- Expand chemical content
- Web-services & Dashboard access





Building DSSTox_v2



DSSTox_v2 Construction

Data source load order:

1) DSSTox_v1 (~22K)
   - ✓ 1:1 CAS-structure mappings
   - ✓ Assign NOCAS_GSID
   - ✓ Related CAS & structure mappings (e.g., NOCAS, mixtures)

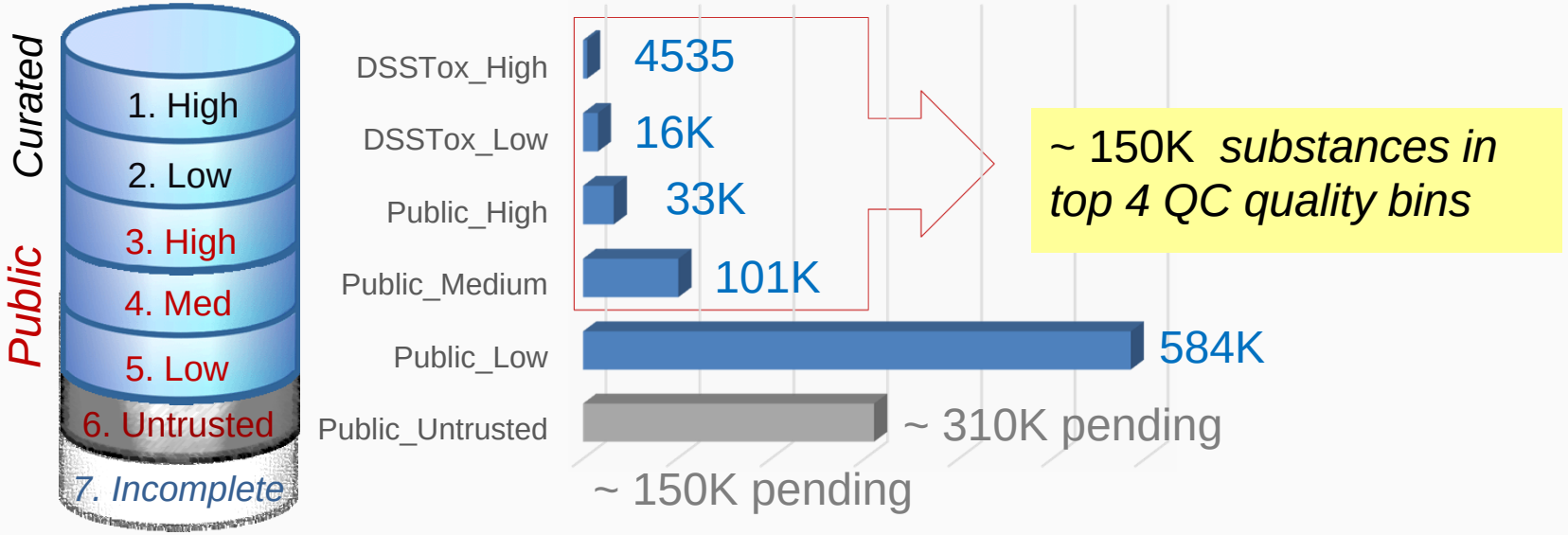
2) EPA SRS (~77K)
   - ✓ *systematic name → structure conversion*
   - ✓ *internal CAS-structure conflicts (12.5%)*
   - ✓ *ChemID conflicts (24% of 30K overlaps)*
   - ✓ *DSSTox conflicts (8% of 6200 overlaps) → queue for curation*

3) ChemID (~400K)
   - ✓ *internal CAS-structure conflicts (4.5%)*
   - ✓ *PubChem conflicts (45% of 225K overlaps) … OUCH!!*
   - ✓ *DSSTox conflicts (11% of 2300 overlaps) → queue for curation*

4) And so on …

DSSTox_v2 Totals

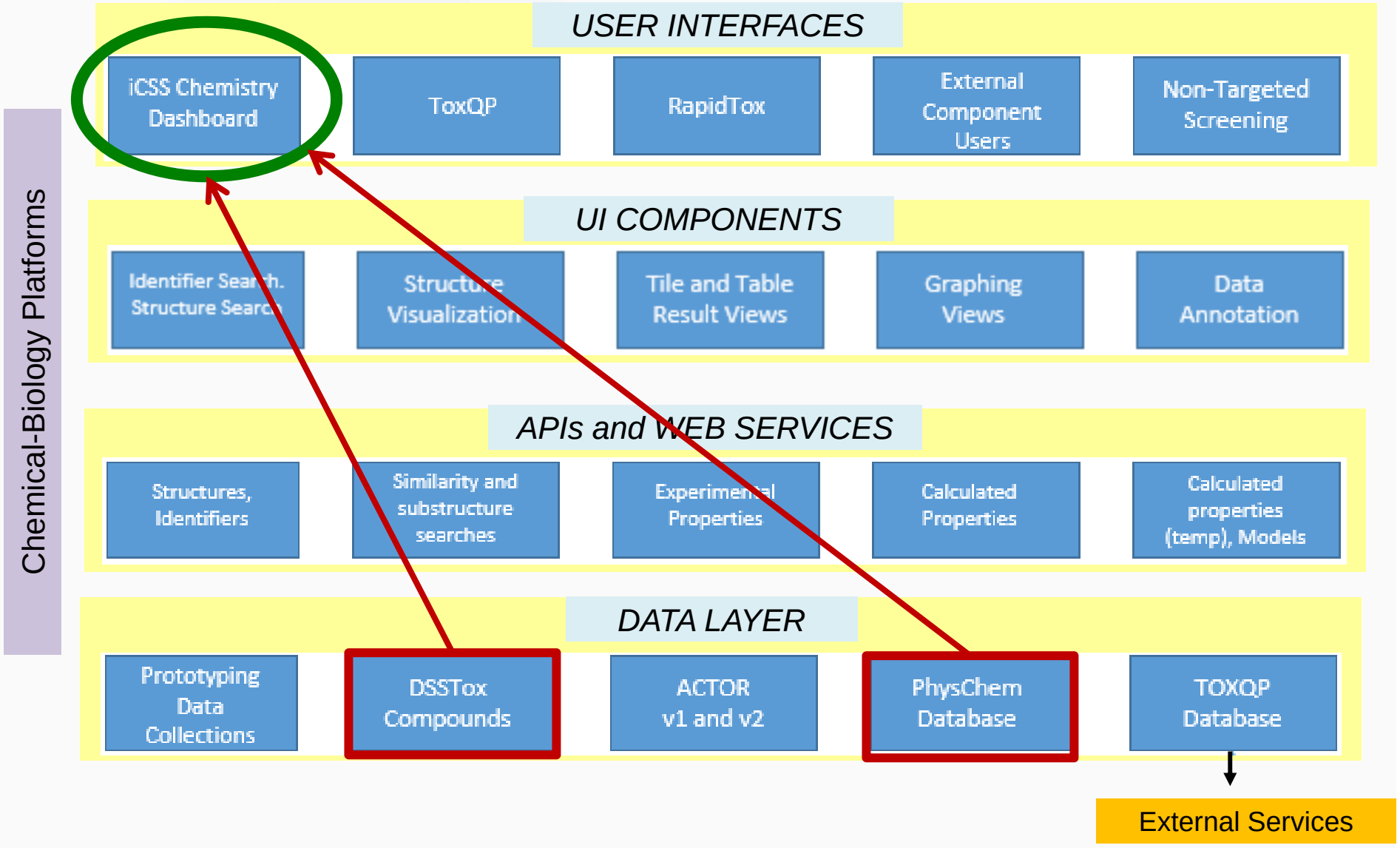
QC Level Totals (12Jun2015)



QC Levels

DSSTox_High:	Hand curated - highest confidence
DSSTox_Low:	Hand curated and confirmed using multiple public sources
Public_High:	Extracted from EPA SRS and confirmed to have no conflicts in ChemID and PubChem
Public_Medium:	Extracted from ChemID and confirmed to have no conflicts in PubChem
Public_Low:	Extracted from ACToR or PubChem (single source)
Public_Untrusted:	Postulated, but found to have conflicts in public sources

Building the Chemistry Software Architecture



Building Predictive Models

NCCT Modeling Activities

- Receptor-mediated activities (e.g., ER, AR)
- Toxicity endpoint QSAR models
- Metabolism prediction
- In vitro-in vivo toxicity extrapolation models (IVIVE)
- Read-across toxicity estimation
- Near and far-field exposure models

EPA-NCCT Collaborations

- Analytical dust and media screening for environmental contaminants
- Biotransformation, transport, fate modeling
- ADME, PBPK models
- Chemical use category modeling
- Green chemistry

Physical chemical properties

Water solubility

MP, BP, Vapor Pressure

LogP

Partition coefficients (air/water/soil)

Building a physical chemical property database

Measured (where available)

- Collect data *(sort wheat from chaff)*
- Determine how to best mesh data together *(structure? CAS? Other?)*
- Check data self-consistency
- Structure validation vs. property value validation *(very different challenges)*

Physical chemical properties

Predicted (where not available)

- Build training/test sets of chemicals with experimental values *(collapse or resolve dups)*
- Apply structure-cleaning workflow to produce "QSAR-ready" structures
- Calculate descriptors
- Apply modeling algorithm *(regression, ML, etc.)*

Store measured & predicted values in relational database, linked to data sources, model details, etc.

Curating structure-linked data

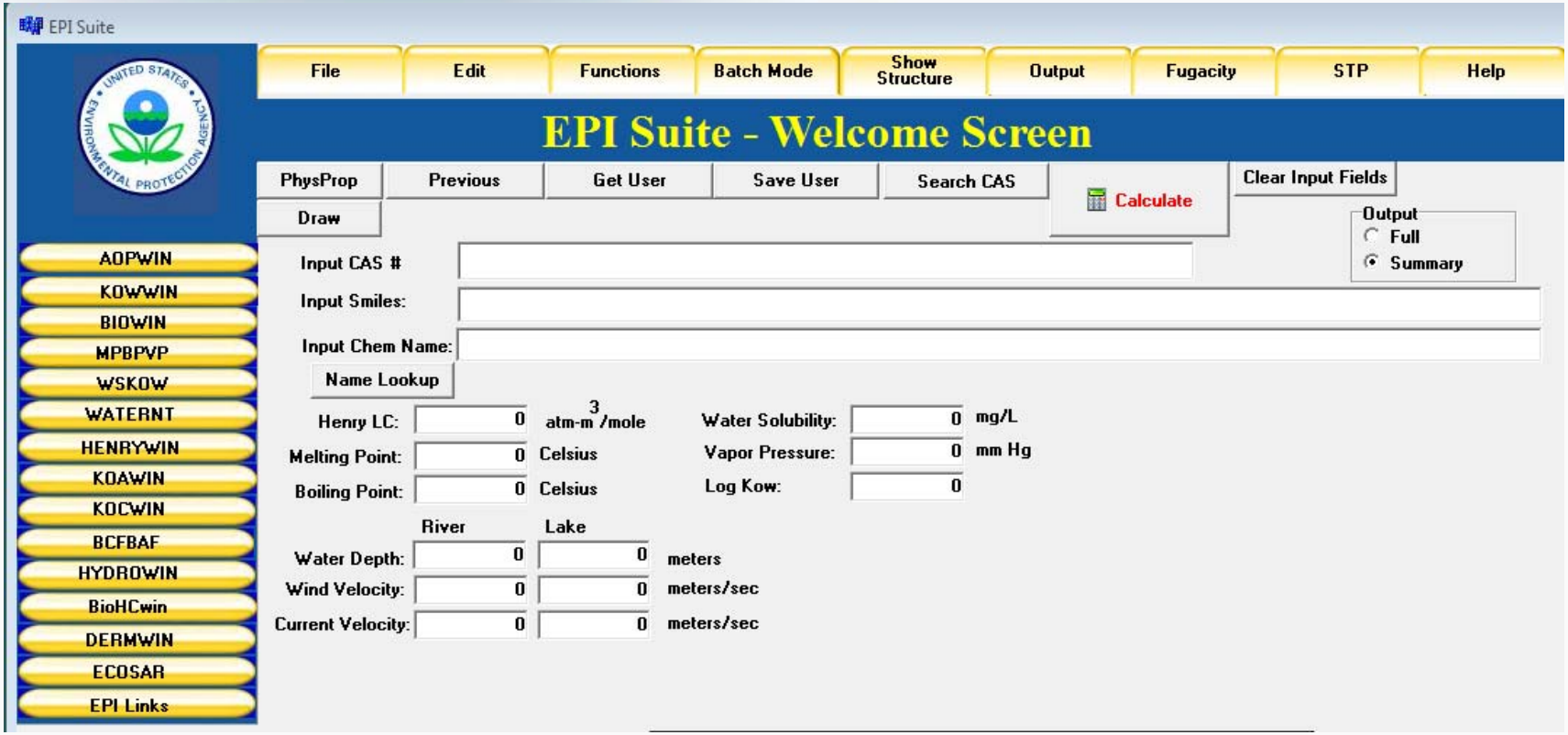
- Establish accurate linkage of measured data to chemical structure
- Published experimental data → chemical IDs {CAS, name, and/or structure}
  - *Fix ID errors*
  - *Resolve ID conflicts*
  - *Normalize, clean structures*
  - *Resolve duplicate mappings*

Lots of experience from building DSSTox!

EPI Suite™ Property Estimation Software

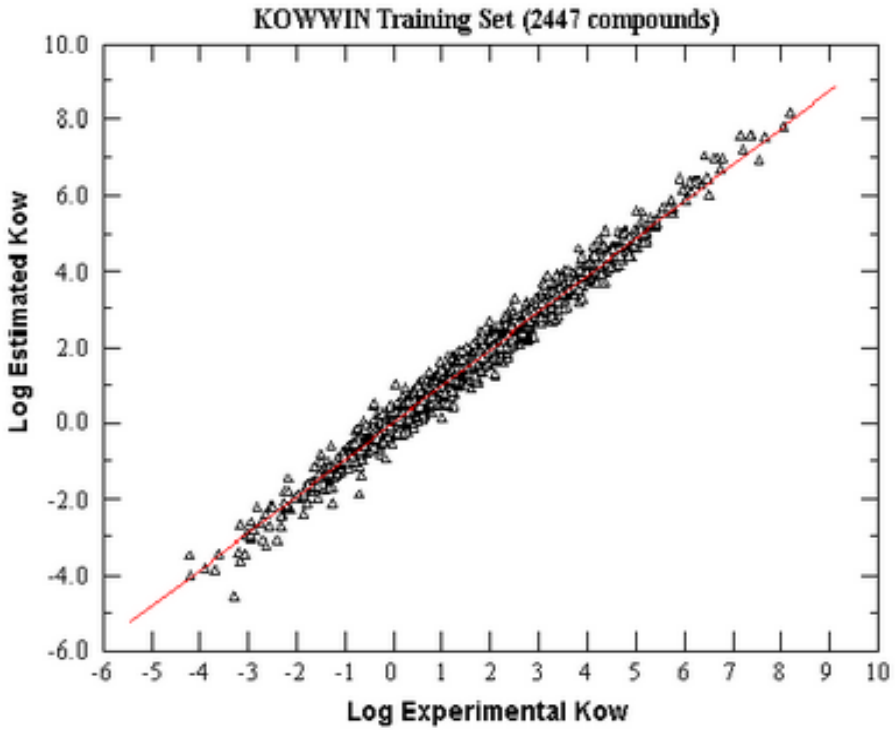
- Developed by EPA's Office of Pollution Prevention & Toxics (w/ Syracuse Research Corp), published initial models over 20 yrs ago, released desktop app in 2000
- Estimates a wide range of physical and environmental properties using QSAR approaches,
  - KOWWIN™, AOPWIN™, HENRYWIN™, MPBPWIN™, BIOWIN™, BioHCwin, KOCWIN™
- and incorporates these into dependent fate and toxicity models,
  - WSKOWWIN™, WATERNT™, BCFBAF™, HYDROWIN™, KOAWIN, AEROWIN™, WVOLWIN™, STPWIN™, LEV3EPI™, ECOSAR™
- Used to fill data gaps in EPA Pre-Manufacture Notification (PMN) chemical submissions, wide usage outside EPA

EPI Suite™ Property Estimation Software



- Downloadable Windows application available at: http://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface
- 14 PhysProp experimental property data sets used in training models can be freely accessed from within app

EPI Suite KOWWIN: 2447 Training Compounds



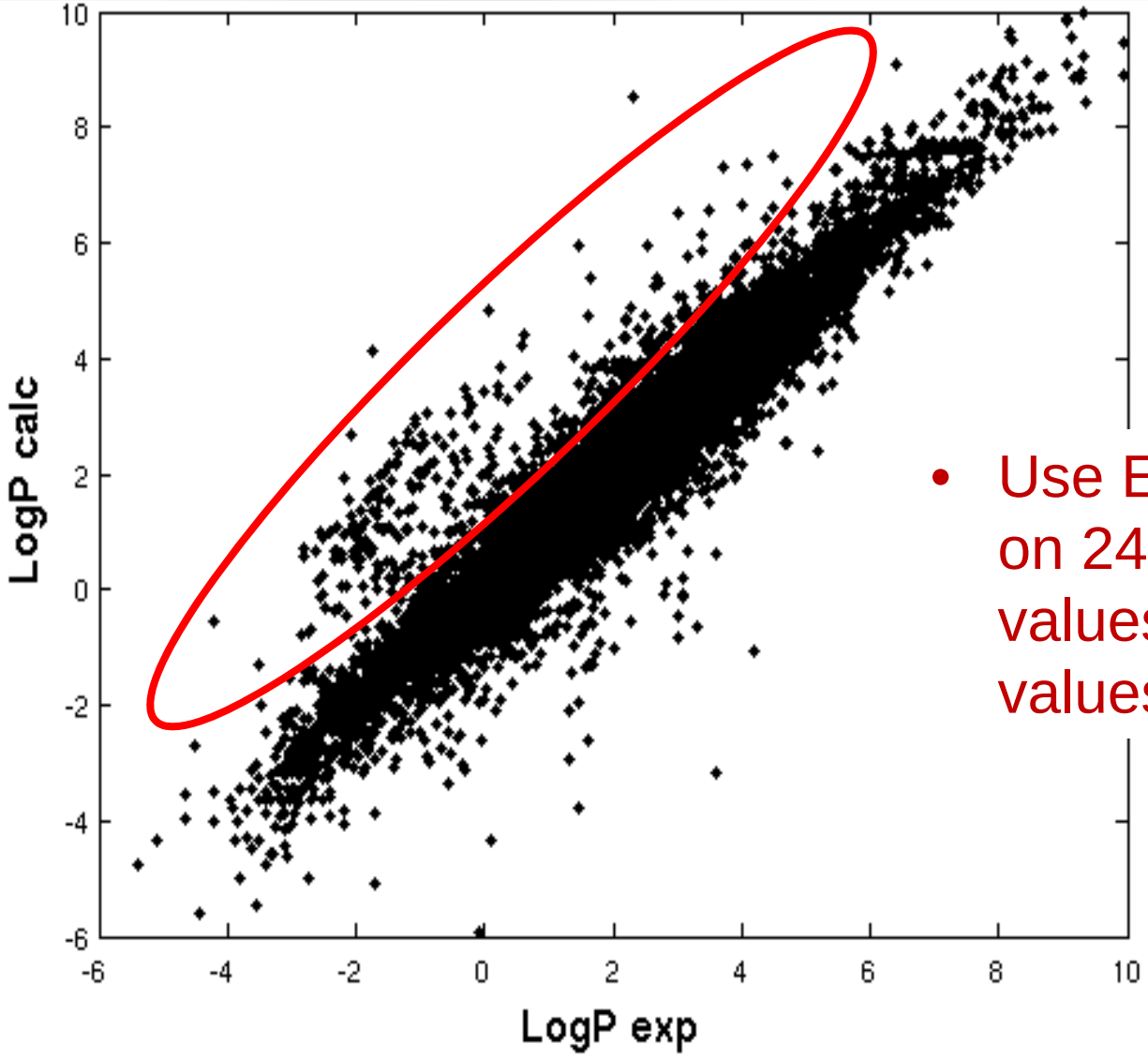
Total Training Set Statistics:

number in dataset	= 2447
correlation coef (r^2)	= 0.982
standard deviation	= 0.217
absolute deviation	= 0.159
avg Molecular Weight	= 199.98

- 2447 chemicals with measured Log Kow values used to train model
- >15K measured Log Kow values available in PhysProp file

Overall Predicted vs. Experimental

(>15k chemicals)



- Use EPI Suite model built on 2447 measured values to predict >15K values in PhysProp file

Why derive new models from EPI Suite PhysProp datasets?

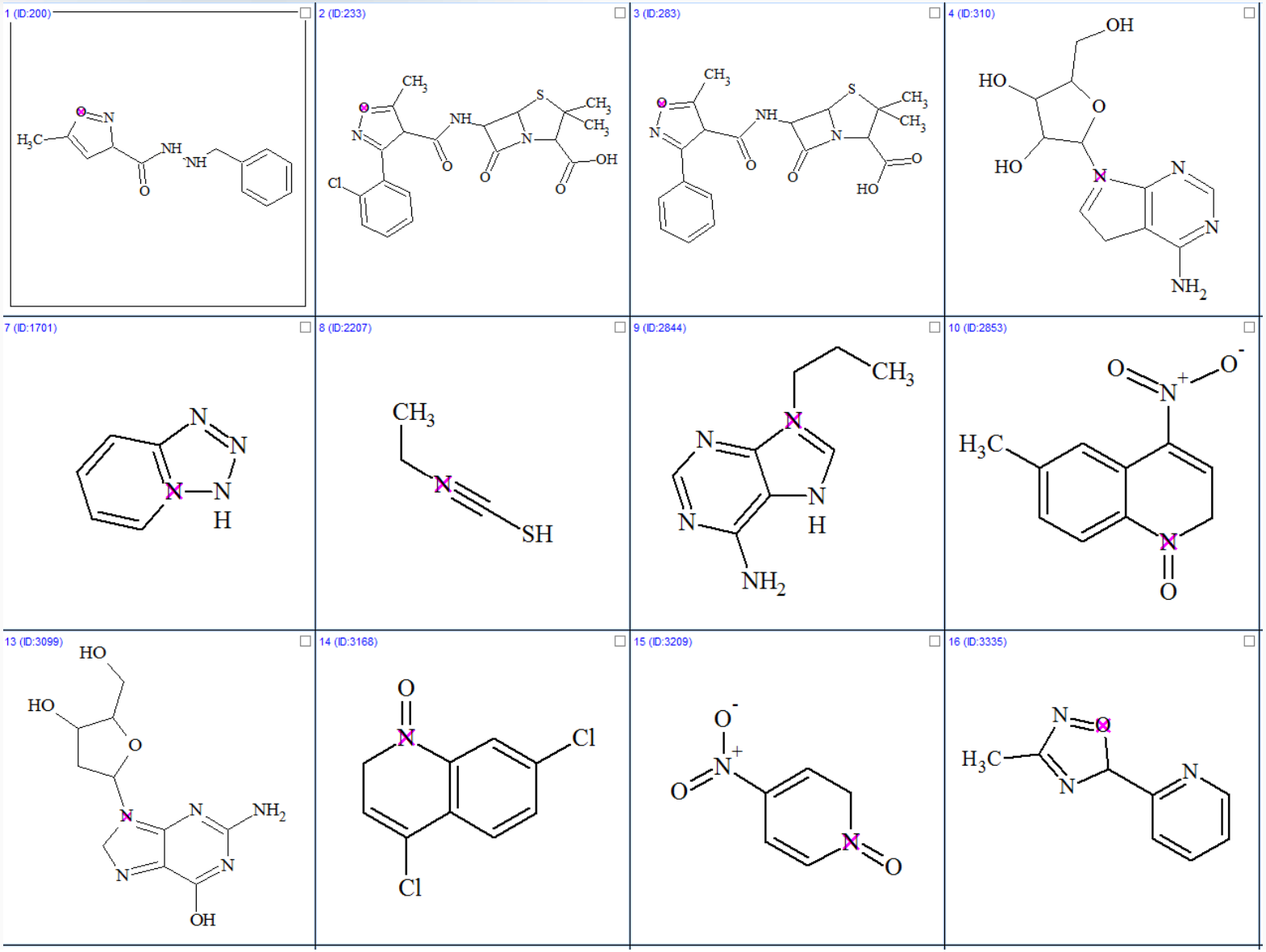
- Significant advances in cheminformatics and modeling approaches since <2000
- Enlarged training sets likely to improve models and expand applicability domain
- But first…

Perhaps we should take a look at the >20 yr old PhysProp datasets that we'll be using to rebuild the models!

Review of PhysProp datasets

- Exported SDF files from EPI Suite application
- Basic manual review searching for errors:
  - hypervalency, charge imbalance, undefined stereo
  - deduplication
  - mismatches between identifiers
    - ✓CAS Numbers not matching structure
    - ✓Names not matching structure
    - ✓Collisions between identifiers

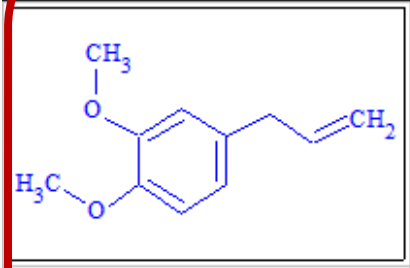
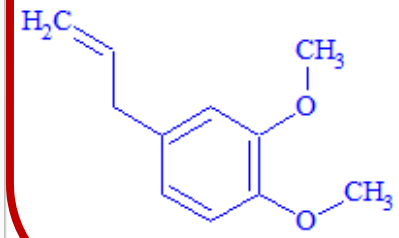
Incorrect valences



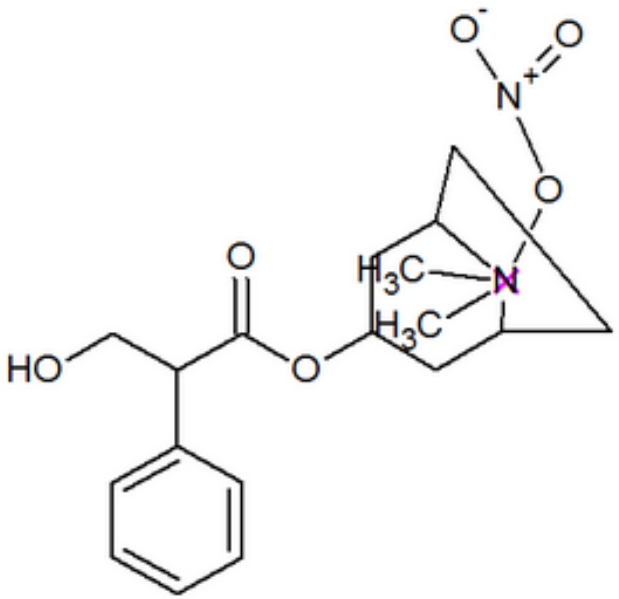
Differences in Values (MP)

Two experimental records:
Same structure

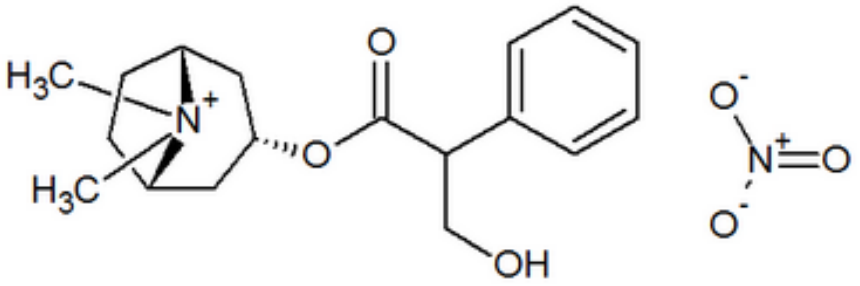
Different CAS, name, experimental & predicted MP

Structure	Formula	FW	CAS	NAME	MP	EstMP	ErrorMP
	$C_{11}H_{14}O_2$	178.2277	000093-15-2	METHYLEUGENOL	-4.000000000000000e+000	3.285000000000000e+001	3.685000000000000e+001
	$C_{11}H_{14}O_2$	178.2277	006380-24-1	4-Allyl-1,2-dimethoxy-benzene	7.000000000000000e+001	3.285000000000000e+001	-3.715000000000000e+001

Covalently bound salt structures



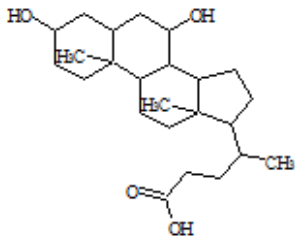
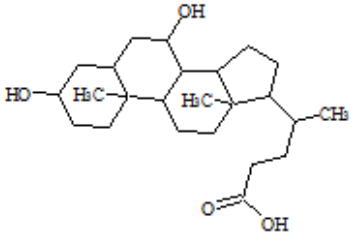
PHYSPROP database



Appropriate Depiction

Collisions in Records

Same structure depictions (Molfiles) - different CAS, different names, AND different SMILES

	$C_{24}H_{40}O_4$	392.5720	000083-49-8	HYODEOXYCHOLIC ACID	3.0800000000000 00e+000
	$C_{24}H_{40}O_4$	392.5720	000128-13-2	URSODEOXYCHOLIC ACID	3.0000000000000 00e+000

Missing or erroneous identifiers

- Many chemical names are truncated
- Many chemicals don't have CAS Numbers

SRC000-02-7	Ethanaminium, N,N,N-trimethyl-2-[(1-oxo-2-propen
SRC000-04-3	Guanidine, N-hydroxy-N''-[4-(methylthio)benzeneme
SRC000-04-4	Hydrazinecarboximidamide, N'-[4-(methylthio)benz
SRC000-04-5	NNN5-TeMe-N-(3FuranMe),ammon Br
SRC000-04-6	Benzenamine, 4-bromo-N,N-bis(2,2,2-trifluoroethy
SRC000-04-7	2-Propenoic acid, 3-(2-chlorophenoxy)-, methyl e
SRC000-05-1	9H-Purine-9-acetaldehyde, a-(1-formyl-2-hydroxye
SRC000-05-2	N1-Pr-N2-CN-N3-Me guanidine
SRC000-05-3	1-(2-OHEt)-2-Me imidazoline HCL

Whoa!! Lots of problems…

15,809 chemicals in KOWWIN data file

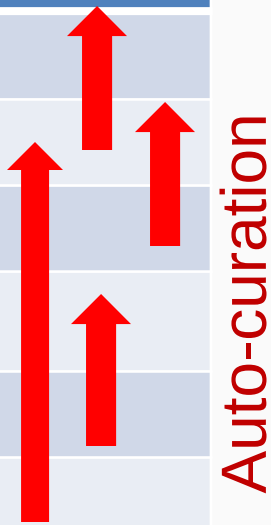
- Invalid CAS Checksum: 3646
- Invalid names: 555
- Invalid SMILES: 133
- Valence errors: 322 Molfile, 3782 SMILES
- Duplicates check:
  - 31 MOLFILE; 626 SMILES; 531 NAMES
- SMILES vs. Molfiles (structure check):
  - 1279 differ in stereochemistry
  - 362 “covalent halogens”
  - 191 differ as tautomers
  - 436 are different compounds

Quality flags added to data: 1-4 Stars

4 levels of consistency possible

- Molblock
- SMILES string
- chemical name (based on ACD/Labs dictionary)
- CAS Number (based on a DSSTox lookup)

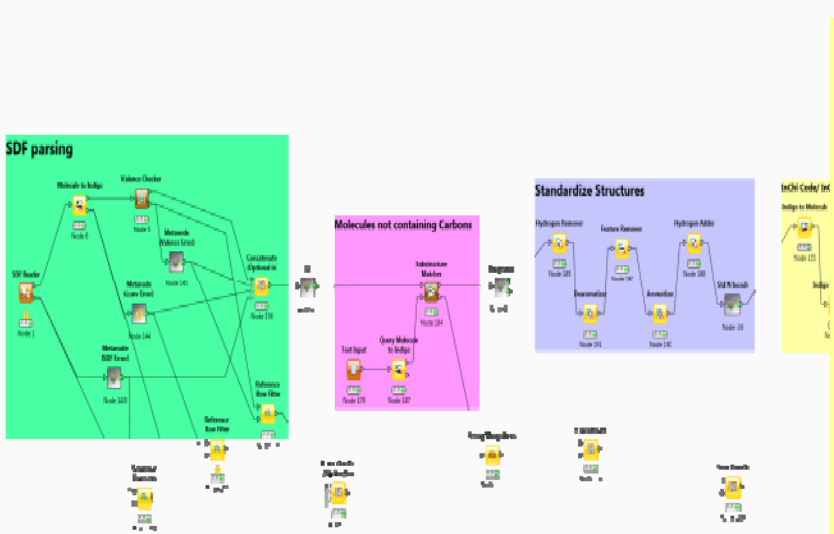
FLAG	Definition
4 Stars ENHANCED	4 levels of consistency with stereo information
4 Stars	4 levels of consistency, stereo ignored
3 Starts Plus	3 of 4 levels consistent; 4th is a tautomer
3 Stars ENHANCED	3 levels of consistency with stereo information
3 Stars	3 levels of consistency, stereo ignored.
2 Stars Plus	2 of 4 levels consistent; 3rd is a tautomer
1 Star	Whatever's left – too many errors to count!



KNIME structure-“cleaning” workflow

https://www.knime.org/knime

- Combine community approaches to structure processing (CERAPP)
- Develop a flexible workflow to be used by EPA and shared publicly
- Process DSSTox files to create “QSAR-ready” structures



- ✓ Parse SDF, remove fragments
- ✓ Explicit hydrogen removed
- ✓ Dearomatization
- ✓ Removal of chirality info, isotopes and pseudo-atoms
- ✓ Aromatization + add explicit hydrogens
- ✓ Standardize Nitro groups
- ✓ Other tautomerize/mesomerization
- ✓ Neutralize (when possible)

Publicly available cheminformatics toolkits in KNIME:

Indigo

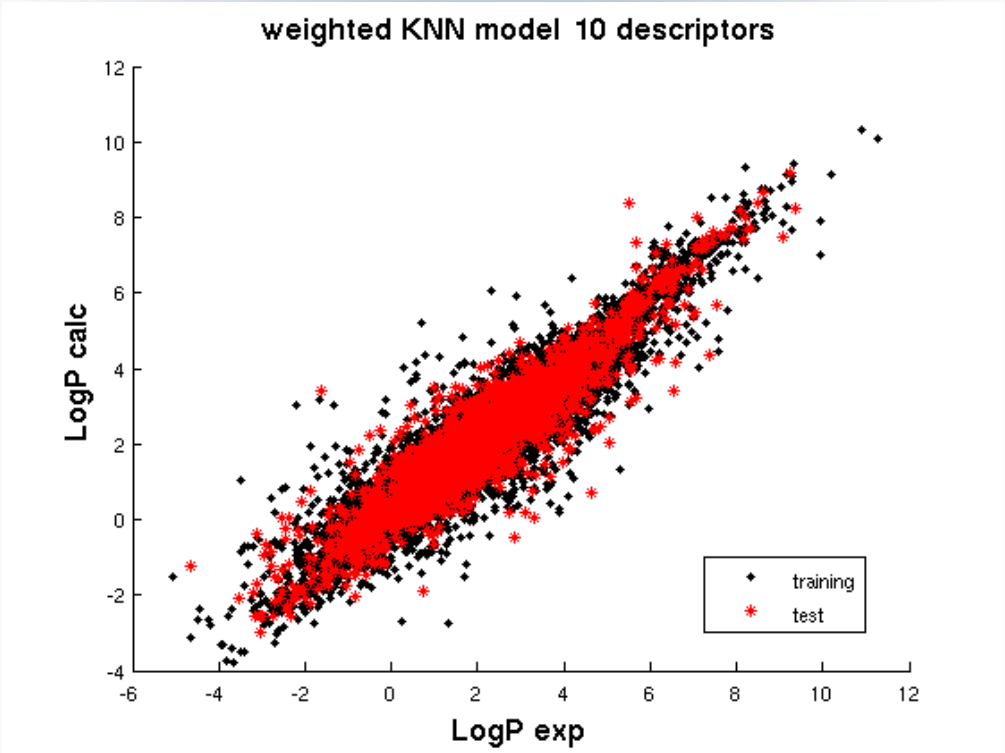
RDKit

Pharmaceutical Data Exploration Laboratory

Building the Models

- QSAR ready forms for modeling – standardize tautomers, remove stereochemistry, no salts
- Remove approx. 1800 chemicals due to poor quality score
- Build using 3 STAR and BETTER chemicals

Rederived model for PhysProp Log Kow (n=14049)



Weighted kNN model, 5-nearest neighbors
Training: 11251 chemicals
Test set: 2798 chemicals

5 fold cross validation:
R2: 0.87
RMSE: 0.67

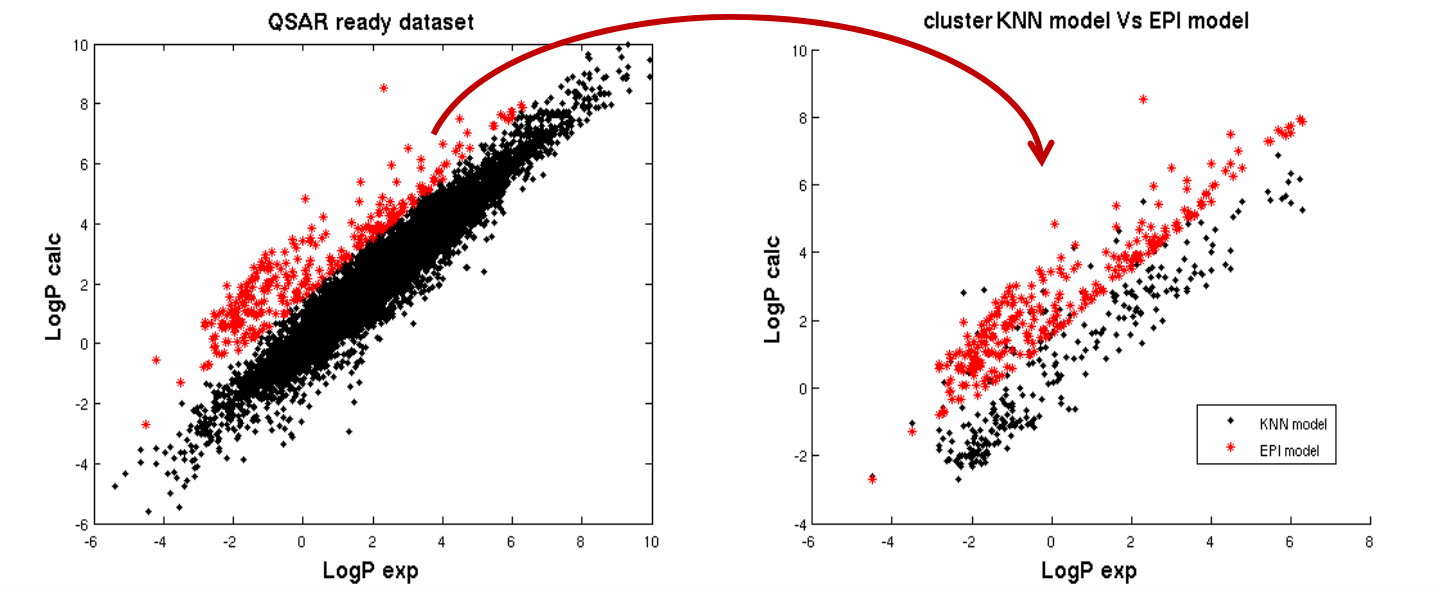
Open source PaDEL descriptors

"'CrippenLogP'" Crippen's LogP
"'GATS2c'" Geary autocorrelation - lag 2 / weighted by charges
"'LipoaffinityIndex'" Lipoaffinity index
"'AATS1p'" Average Broto-Moreau autocorrelation - lag 1 / weighted by polarizabilities
"'ATSC1i'" Centered Broto-Moreau autocorrelation - lag 1 / weighted by 1st ionization potential
"'ETA_EtaP'" Composite index Eta relative to molecular size
"'MLFER_S'" Combined dipolarity/polarizability
"'nN'" Number of nitrogen atoms
"'ETA_Beta'" A measure of electronic features of the molecule
salt _ index salt info

Remember the original EPI Suite outlier scatter?

- Applicability domain (AD) of original EPI Suite LogKow Model improved in updated PhysProp model

EPISuite Predicted vs. Experimental **EPISuite vs kNN Comparison of Cluster**



** 280 cmpd cluster outside AD of EPI Suite → no longer outliers in new model*

Updated PhysProp results

- Performed curation and cleaning of all 14 PhysProp measured property datasets
- Rederived models using modern ML methods & open source descriptors provide significantly improved prediction accuracy over older EPI Suite models
- Predictions generated for entire (>700K) DSSTox database, stored in new property database
- Measured & predicted properties to be surfaced in first release of iCSS Chemistry Dashboard (April 2016)

→ Global statistics insensitive to local curation improvements in >15K training set, BUT exceedingly important when surfacing measured properties in support of individual prediction values

iCSS Chemistry Dashboard Releasing April 2016

- PHASE 1 Delivery – Web interface supporting CSS research
  - access to DSSTox content: >700,000 chemicals
  - access to experimental property data
- Initial set of models based on reanalysis of cleaned, curated EPI Suite PHYSPROP datasets – logP, BP, MP, Wsol etc.

DEVELOPMENT

iCSS Chemistry Dashboard Releasing in April 2016

EPA United States Environmental Protection Agency

EPA United States Environmental Protection Agency

Search Chemical

DEVELOPMENT

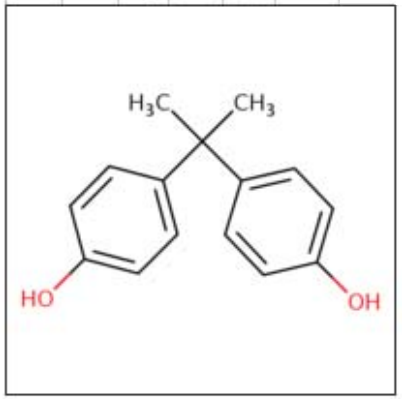
Bisphenol A

80-05-7 DTXSID7020182

Formerly DSSTox GSID, new unique public substance ID, essential for RDF/semantic web applications

Searched by Synonym: Found 1 result for 'bisphenol A'.

2D 3D



Intrinsic Properties

Molecular Formula: C15H16O2 Search in DSSTox

Average Mass: 228.291 g/mol

Monoisotopic Mass: 228.11503 g/mol

Structural Identifiers

Citation

Chemical Properties | External Links | Synonyms | PubChem Biological Activities | PubChem Articles | PubChem Patents | Comments

CSV Excel

Property	Average (Exp.)	Range (Exp.)	Average (Pred.)	Range (Pred.)
Solubility	0.001 (1)	0.0005257 to 0.0005257	0.38 (2)	0.003675 to 0.7565
Melting Point	154.929 (7)	153.0 to 158.0	144.033 (3)	131.8 to 158.0
Boiling Point	200.0 (1)	200.0 to 200.0	348.95 (2)	334.4 to 363.5
LogP	3.357 (3)	3.32 to 3.431	3.524 (3)	3.205 to 3.727
Atmospheric Hydroxylation Rate	N/A	N/A	0.0 (1)	4.237e-11 to 4.237e-11

About Contact ACToR DSSTox Privacy Accessibility Help

iCSS Chemistry Dashboard Releasing in April 2016

Chemical Properties: Melting Point

	Average	Range
Experimental	154.929 (7)	153.0 to 158.0
Predicted	144.033 (3)	131.8 to 158.0

- Original raw source result view
- Users can submit comments

CSV | Excel

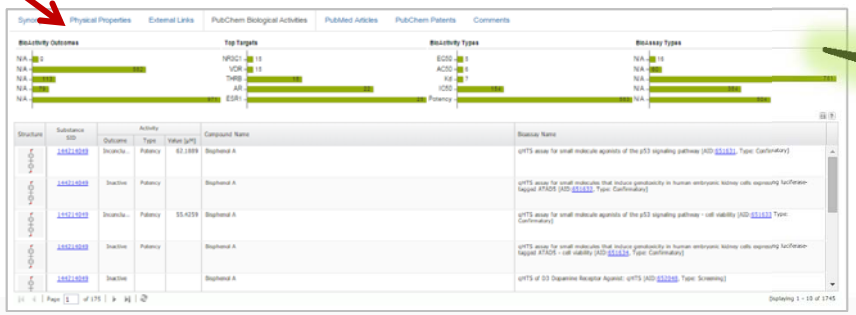
Property	Raw Result	Mean Result	Minimum Result	Maximum Result	Result Unit	Result Type	Source
Estimated MP (oC)	131.76	131.8	131.8	131.8	°C	predicted	EPI SUITE
Melting Point	153-158 °C	155.5	153.0	158.0	°C	experimental	Alfa Aesar
Melting Point	154-157 °C	155.5	154.0	157.0	°C	experimental	Merck Millipore
Melting Point	153-158 °C	155.5	153.0	158.0	°C	experimental	Alfa Aesar
Melting Point	155-158 °C	156.5	155.0	158.0	°C	predicted	J and K Scientifi
Melting Point	156 °C	156.0	156.0	156.0	°C	experimental	TCI
Melting	153 °C	153.0	153.0	153.0	°C	experimental	Jean-Claude Br

iCSS Chemistry Dashboard Releasing in April 2016

- Links to external resources: EPA, NIH, property predictors

Chemical Properties | External Links | Synonyms | PubChem Biological Activities | PubChem Articles | PubChem Patents

General	Toxicology	Publications	Biochemistry	Prediction
ToxCast Dashboard 2	ACTOR	Toxline	CCRIS	Chemicalize
EPA Substance Registry Service	DrugPortal	Environmental Health Perspectives		
NIST Chemistry Webbook		NIEHS		
eChemPortal		National Toxicology Program		
Household Products Database		Google Books		
HSDB				
PubChem				
Chemspider				
EDSP Dashboard				
CPCat				
DrugBank				
HMDB				



e.g., ChemAxon's "Chemicalize" web-service

e.g., PubChem's bioactivity summary

Linkage to External Predictor, e.g. Chemicalize

Chemicalize.org (powered by ChemAxon)

www.chemicalize.org/structure/#!m

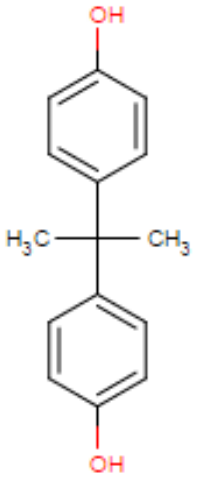
chemicalize.org beta
by ChemAxon

Properties Viewer Webpage Viewer Chem Search Web Search Doc

CC(C)(C1=CC=C(O)C=C1)C1=CC=C(O)C=C1

Molecule

Molecule



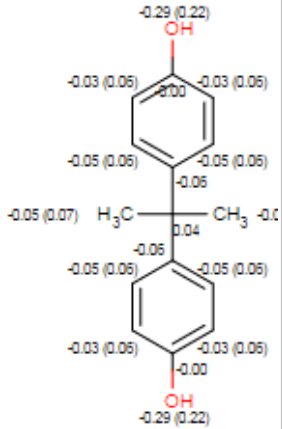
Names and identifiers

Common names: 4,4'-bisphenol A, 4,4'-bisphenol A, diphenylolprop
bisphenol A, bisphenol-A, bisphenol A, 4,4' isopropylidinediphenol,
IUPAC: 4-[2-(4-hydroxyphenyl)propan-2-yl]phenol

Smiles: CC(C)(C1=CC=C(O)C=C1)C1=CC=C(O)C=C1
InChI: 1S/C15H16O2/c1-15(2,11-3-7-13(16)8-4-11)12-5-9-14(17)10-
InChI key: IISBACLAFKSPIT-UHFFFAOYSA-N
CAS: 27100-33-0, 80-05-7, 137885-53-1, 27360-89-0, 28106-82-3,

Charge

Atomic Charges



Elemental Analysis

Formula: $C_{15}H_{16}O_2$
Isotope formula: C15H16O2
Composition: C (78.92%), H (7.06%), O (14.02%)
Isotope composition: C (78.92%), H (7.06%), O (14.02%)
Mass: 228.2863
Exact mass: 228.115029756

Webpages

Sulfone
29 Apr 2015 - original page
wn.com/Sulfone

Microsoft Word - NLPFIN New.doc
27 Mar 2014 - original page
reach-clp-helpdesk.de/de/Downloads/NLP-Liste.pdf?__blob=pul

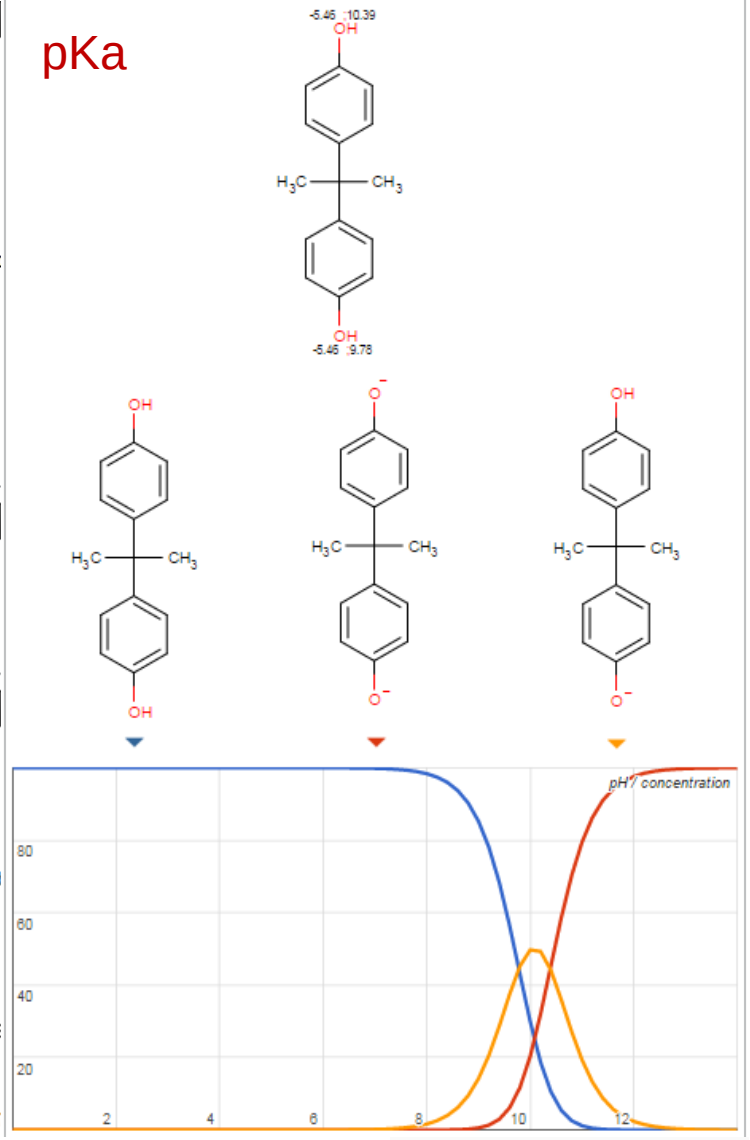
No title
17 May 2014

No title
17 Oct 2012 - original page
patbase.com/getimg/prt_text.asp?id=1551799&pn=EP1248760E

Breast Neoplasms - Genes | CTD
20 Nov 2015 - original page
ctdbase.org/detail.go;jsessionid=E301927D5D648246D5BAA72

pKa

pKa

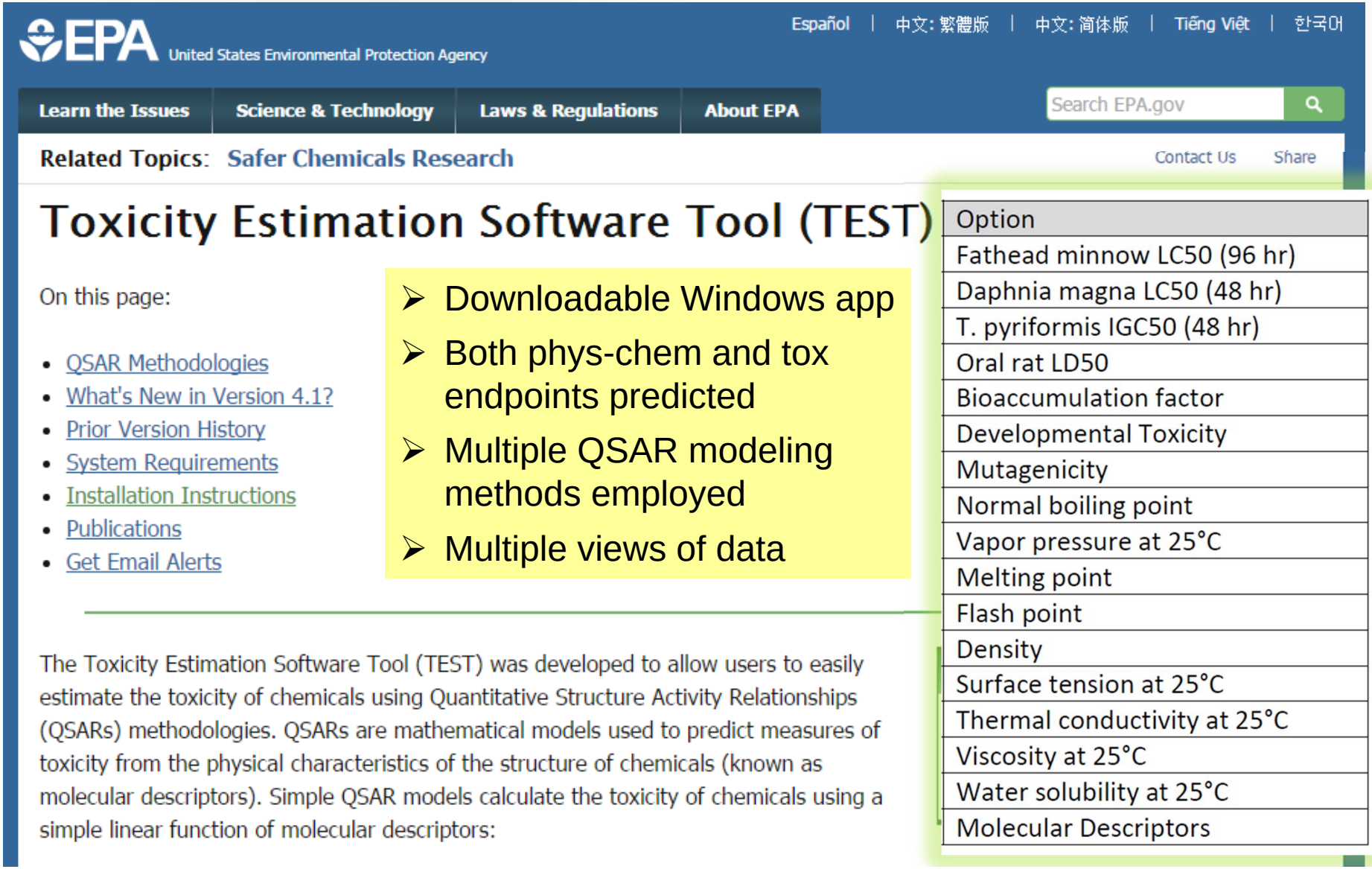


Working to integrate other EPA predictors

- ExpoCast
  - ➢ near & far-field exposure models
- Environmental Fate Simulator (EFS)
  - ➢ air/soil/water distribution, biotransformation
- CERAPP
  - ➢ estrogen receptor activity QSAR model
- T.E.S.T *(in progress)*
  - ➢ phys-chem properties & toxicity endpoints

e.g., T.E.S.T.

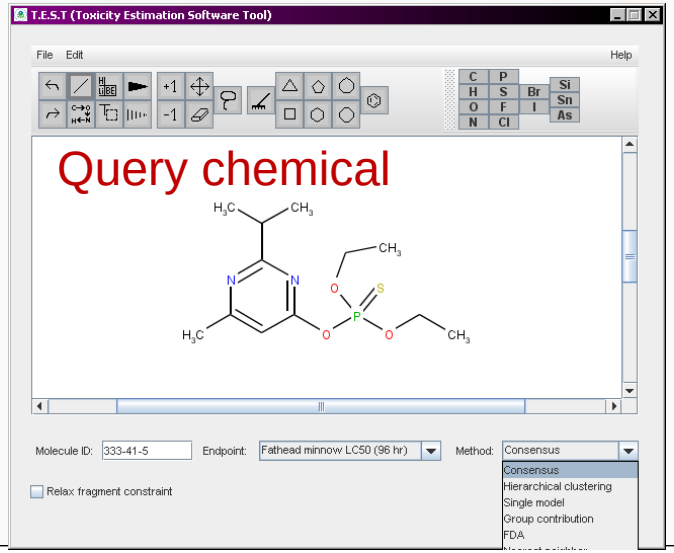
http://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test



- Downloadable Windows app
- Both phys-chem and tox endpoints predicted
- Multiple QSAR modeling methods employed
- Multiple views of data

Option
Fathead minnow LC50 (96 hr)
Daphnia magna LC50 (48 hr)
T. pyriformis IGC50 (48 hr)
Oral rat LD50
Bioaccumulation factor
Developmental Toxicity
Mutagenicity
Normal boiling point
Vapor pressure at 25°C
Melting point
Flash point
Density
Surface tension at 25°C
Thermal conductivity at 25°C
Viscosity at 25°C
Water solubility at 25°C
Molecular Descriptors

T.E.S.T. QSAR Model Prediction Views



Model # 1296

Toxicity prediction results for 333-41-5 for Hierarchical clustering method

Prediction results

Endpoint	Experimental value CAS: 333-41-5 Source: ECOTOX	Predicted value[a]	Prediction interval
Fathead minnow LC_{50} (96 hr) -Log(mol/L)	4.81	5.39	4.54 ≤ Tox ≤ 6.24
Fathead minnow LC_{50} (96 hr) mg/L	4.70	1.23	0.17 ≤ Tox ≤ 8.71

[a]Note: the test chemical was present in the external test set.

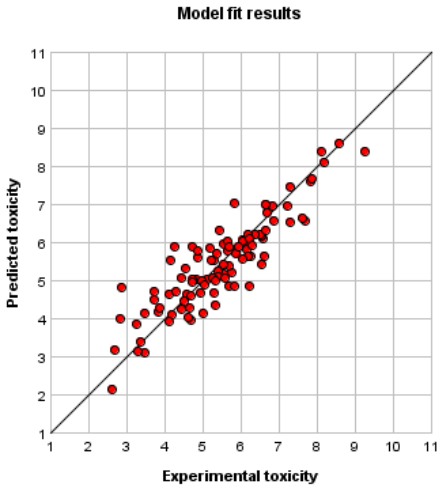
Prediction results

Cluster model predictions and statistics

Cluster model	Test chemical descriptor values	Prediction interval -Log(mol/L)	r^2	q^2	#chemicals
1296	Descriptors	6.010 ± 1.136	0.793	0.733	101
1300	Descriptors	5.458 ± 1.312	0.729	0.645	111
1301	Descriptors	5.136 ± 1.169	0.747	0.718	294
1302	Descriptors	4.922 ± 1.182	0.774	0.751	641

Model statistics

Parameter	Value
Endpoint	Fathead minnow LC_{50} (96 hr)
r^2	0.793
q^2	0.733
#chemicals	101
Model	Model # 1296

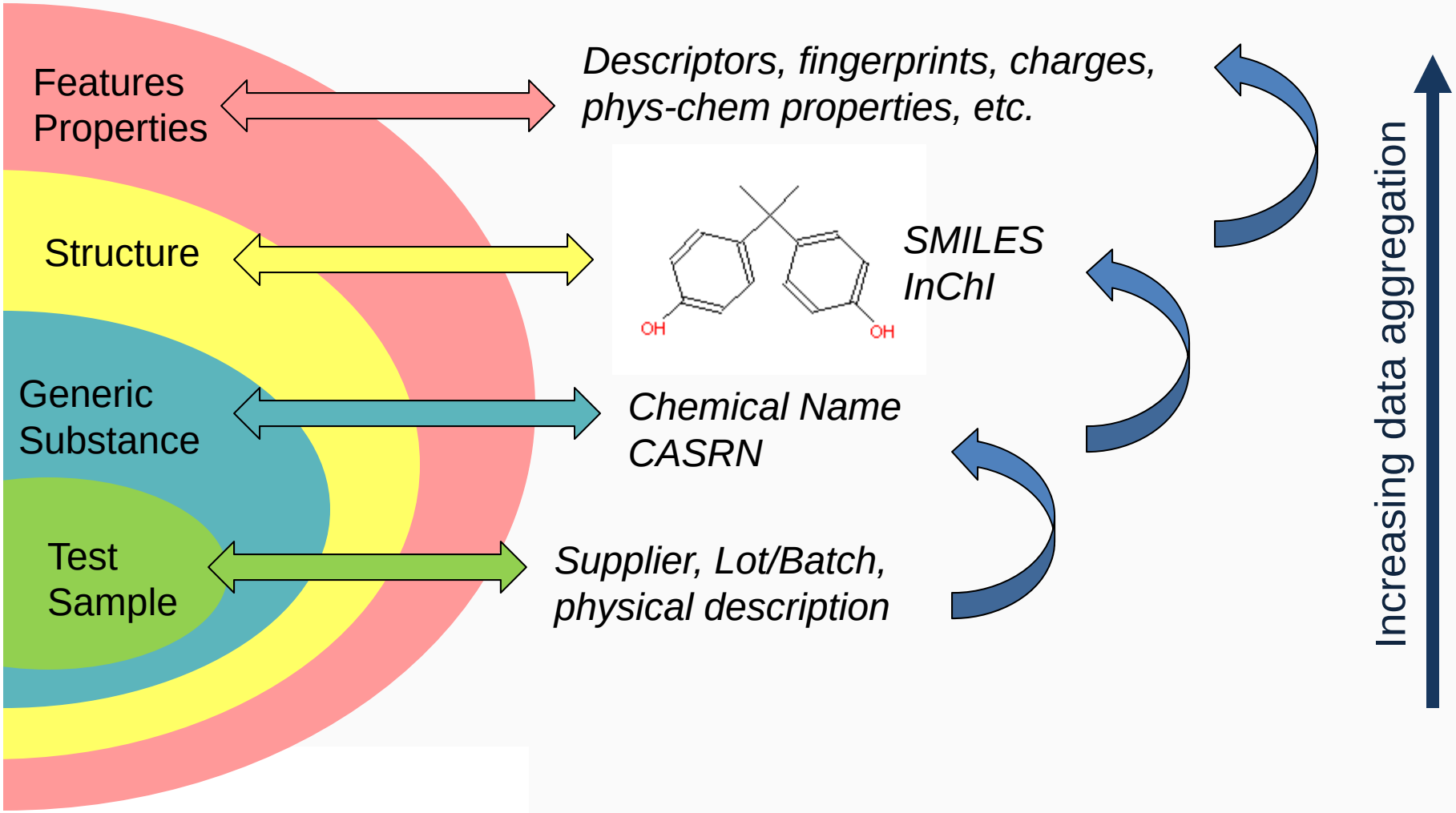


Cluster models with violated constraints

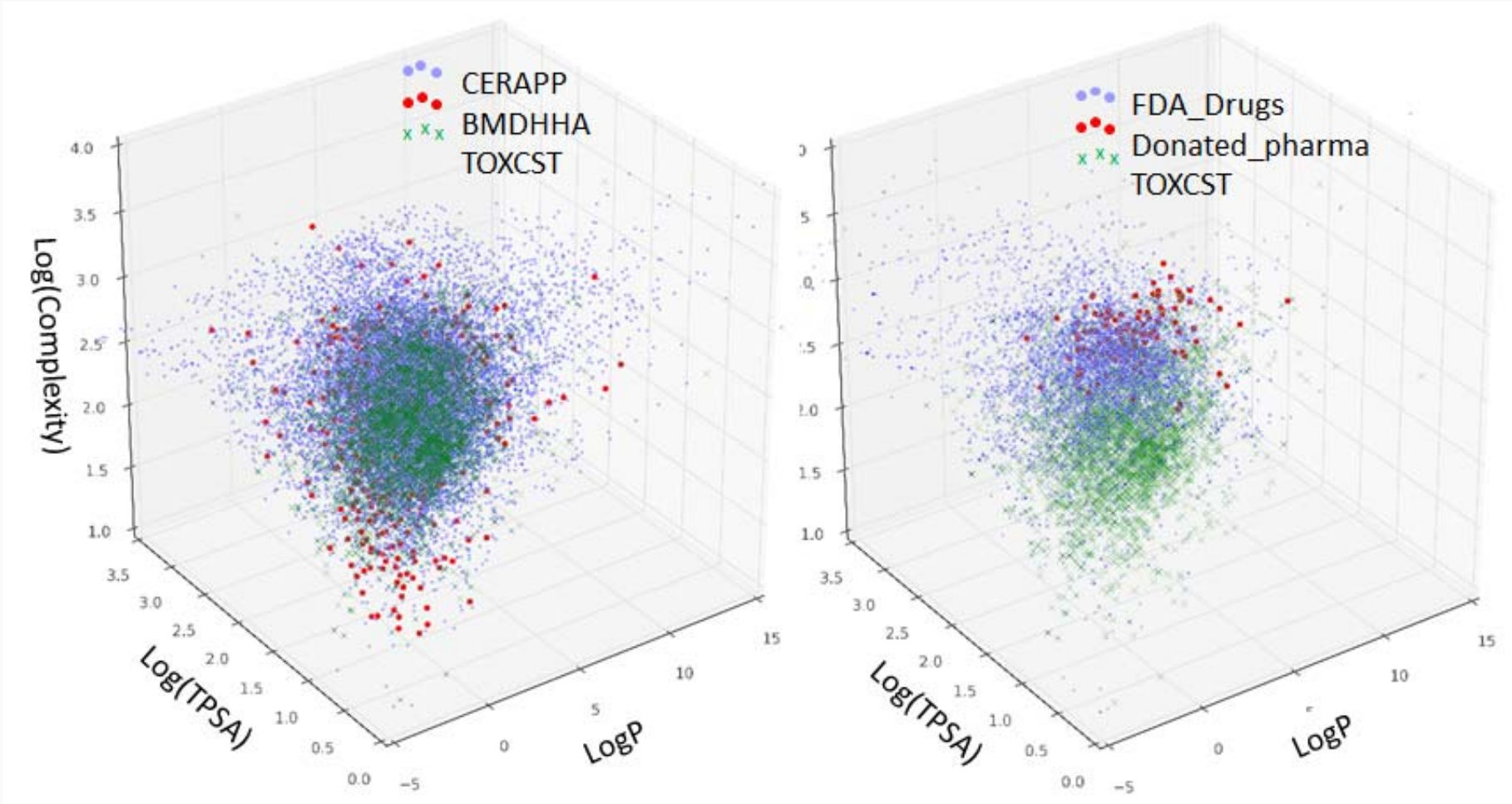
Cluster Model	r^2	q^2	# chemicals	Message
1121	0.810	0.576	10	Rmax constraint not met
1209	0.799	0.574	11	Fragment constraint not met
1247	0.919	0.647	20	Fragment constraint not met
1264	0.869	0.781	22	Fragment constraint not met
1268	0.675	0.553	24	Fragment constraint not met

Descriptor values for test chemical

Quality review required at all levels of chemical "Representations"



Comparing ToxCast to various inventories based on computed properties



In conclusion…

- iCSS Chemistry Dashboard will provide public access to data & services focused on chemicals of interest to EPA
- Initially serve up results for measured & pre-predicted physical chemical property data
- Chemical-Data consistency quality flags for DSSTox and property data add value to public domain data, BUT … *chemical-data linkages require curation/validation!!*
- All data and models will be available as OPEN DATA and OPEN CODE

Stay tuned!!

Acknowledgements

NCCT Contributors to the iCSS Chemistry Dashboard

- Jeff Edwards
- Jeremy Fitzpatrick
- Jordan Foster
- Jason Harris
- Dave Lyons
- Kamel Mansouri
- Aria Smith
- Jennifer Smith
- Indira Thillainadarajah
- Chris Grulke
- Antony Williams