EPA
United States
Environmental Protection
Agency

ToxValDB: Compiling Publicly Available In Vivo Toxicity Data

Richard Judson
U.S. EPA, National Center for Computational Toxicology
Office of Research and Development

Comptox Communities of Practice

December 20, 2018

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

Project Drivers for ToxValDB

- RapidTox: Can we do a rapid, screening-level risk assessment on a chemical (hours to days)?
  - Need access to all available *in vivo* studies or NAM equivalent (New Approach Methods)
  - Example: chemicals found at Superfund sites without RSL values
- Prioritization: How do we select the next chemical to assess?
  - Applications at EPA and other government organizations
- Modeling – Predicting toxicity of data-poor chemicals
  - Most chemicals will never be tested in vivo, but models can be built using existing data
  - QSAR, QBAR

ToxValDB aims to meet these needs

- As many *in vivo* studies as possible

- Focus initially on quantitative values (e.g. NOEL/LOEL)

- Capture key study parameters
  - Study type, exposure route, duration, species, sex, …
  - Where possible, capture critical effects and other information
  - Provide links to original study documents where possible

- Make accessible on-line

Key Data Sources

- ATSDR – US CDC risk assessments
- COSMOS – FDA, cosmetics and food ingredients
- California EPA / OEHHA - Human health benchmarks
- DOD – Military Exposure Guidelines
- DOE – Ecotoxicology risk assessments
- ECHA / REACH – industrial chemicals, human and eco
- EFSA – food additives , human and eco
- EPA ECOTOX – ORD/ MED, pesticides + others
- EPA HEAST - EPA risk assessment values
- EPA HPVIS – OPPT, industrial chemicals
- EPA IRIS - human health risk assessments
- EPA OPP – Pesticide risk assessments
- EPA OW - drinking water standards
- EPA PPRTV DB - 2 versions, NCEA and ORNL, human health risk assessments
- EPA TEST - acute toxicity values from the literature
- EPA ToxRefDB – OPP, pesticidal actives mostly, some literature data, mammalian studies
- HAWC – public studies entered into HAWC from multiple projects
- HESS – Japan, rat subchronic studies on industrial chemicals
- Health Canada - human health values
- WHO IPCS - pesticide risk values
- Wignall – IRIS and literature studies, with BMD values added

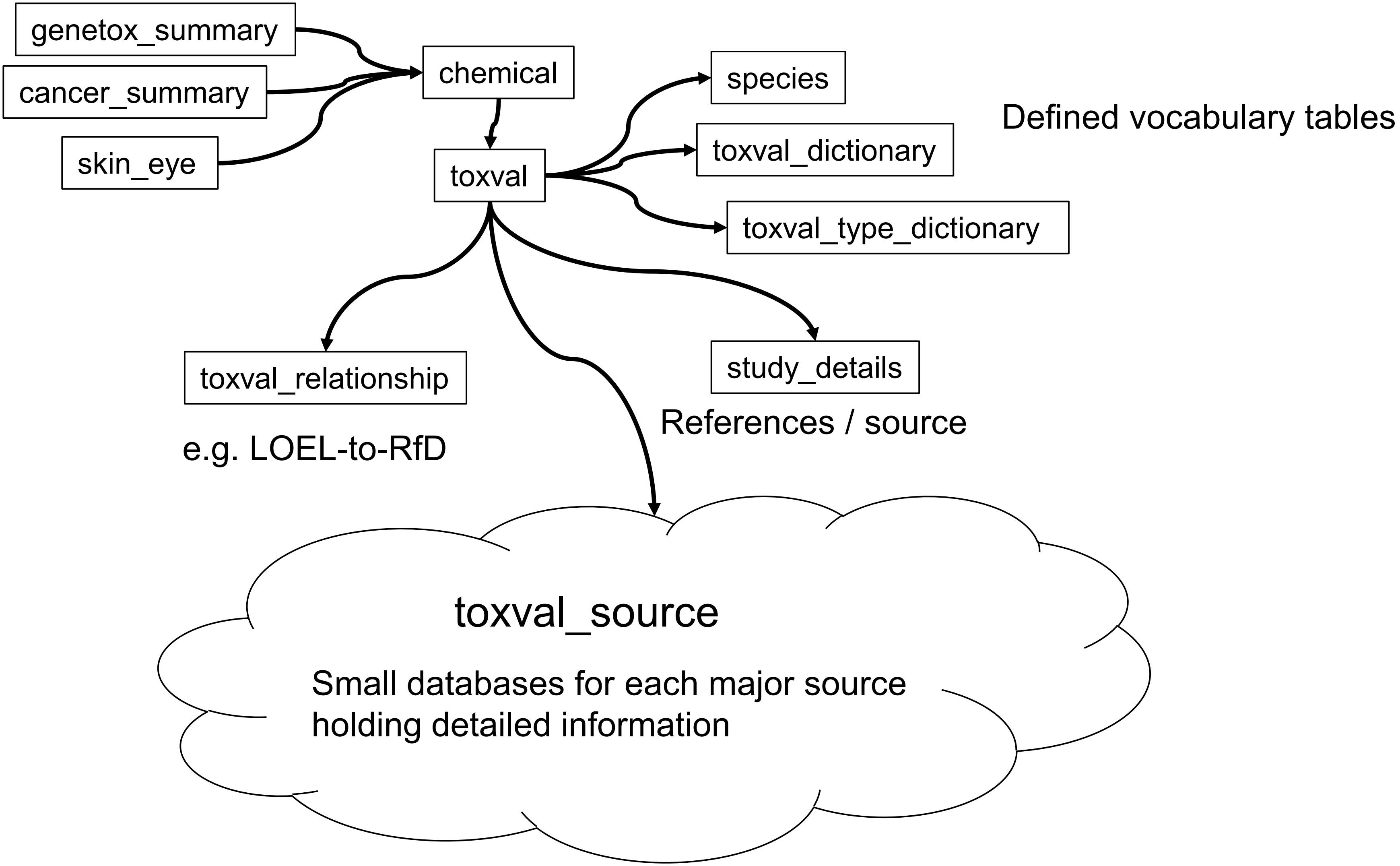
What is a “toxicity value”

- In ToxValDB, a “toxicity value” is a generic name for any quantitative measure
  - LOEL / NOEL .. LOEC / NOEC
  - LOAEL / NOAEL … LOAEC / NOAEC
  - LD50 / LC50
  - BMD / BMC
  - RfD / RfC
  - AEGL, MRL, REL, MEG
  - Cancer slope factor, unit risk
  - Screening level, exposure limits
  - …

Why this is hard …

- Every data source structures the data in a different way, uses different terminology, different units, …
- Curation – automated first
  - Convert to a common vocabulary …
  - Chemical – map to unique DSSTox ID
  - Study type (called "risk assessment class")
  - Value type (called "toxval_type")
  - Value units (convert to a few standard ones: mg/kg-day, mg/L, …)
  - Species, strain, sex
  - Study duration and units
  - Exposure route and method
  - Study year
  - Study information – journal, year, PMID, URL
- Manual curation process being designed

Database structure summary



genetox_summary

cancer_summary

skin_eye

chemical

toxval

species

toxval_dictionary

toxval_type_dictionary

Defined vocabulary tables

toxval_relationship

e.g. LOEL-to-RfD

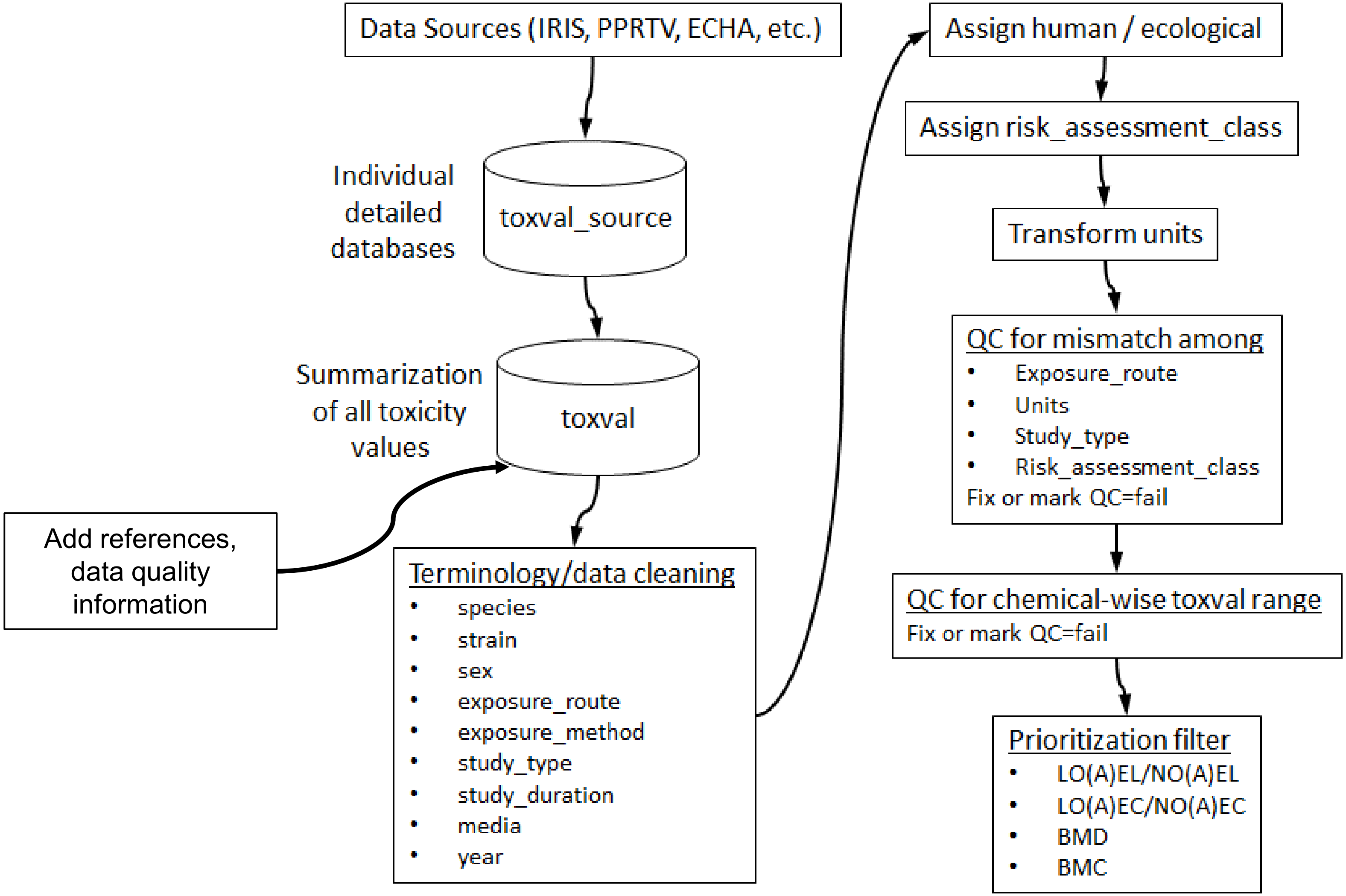
study_details

References / source

toxval_source

Small databases for each major source holding detailed information

ToxValDB Cleaning Process



Data Sources (IRIS, PPRTV, ECHA, etc.)

Individual detailed databases

toxval_source

Summarization of all toxicity values

toxval

Add references, data quality information

Terminology/data cleaning
- species
- strain
- sex
- exposure_route
- exposure_method
- study_type
- study_duration
- media
- year

Assign human / ecological

Assign risk_assessment_class

Transform units

QC for mismatch among
- Exposure_route
- Units
- Study_type
- Risk_assessment_class

Fix or mark QC=fail

QC for chemical-wise toxval range

Fix or mark QC=fail

Prioritization filter
- LO(A)EL/NO(A)EL
- LO(A)EC/NO(A)EC
- BMD
- BMC

Cancer and genotoxicity values being handled separately

- Cancer:
  - Data from IARC, IRIS, NTP, OPP, PPRTV, CalEPA, Health Canada, NIOSH
  - Chemicals can have “cancer classifications” (e.g. “probable”, “possible”, “likely” human carcinogen)
  - Chemicals can also have individual study data
    - Cancer-related critical effects
    - Cancer tox_values (cancer slope factors, unit risk values)
- Genotoxicity
  - Data from COSMOS, ECHA, NLM TOXNET, TEST
  - Sources use different terminology, so all test descriptions were mapped to a common set of terms

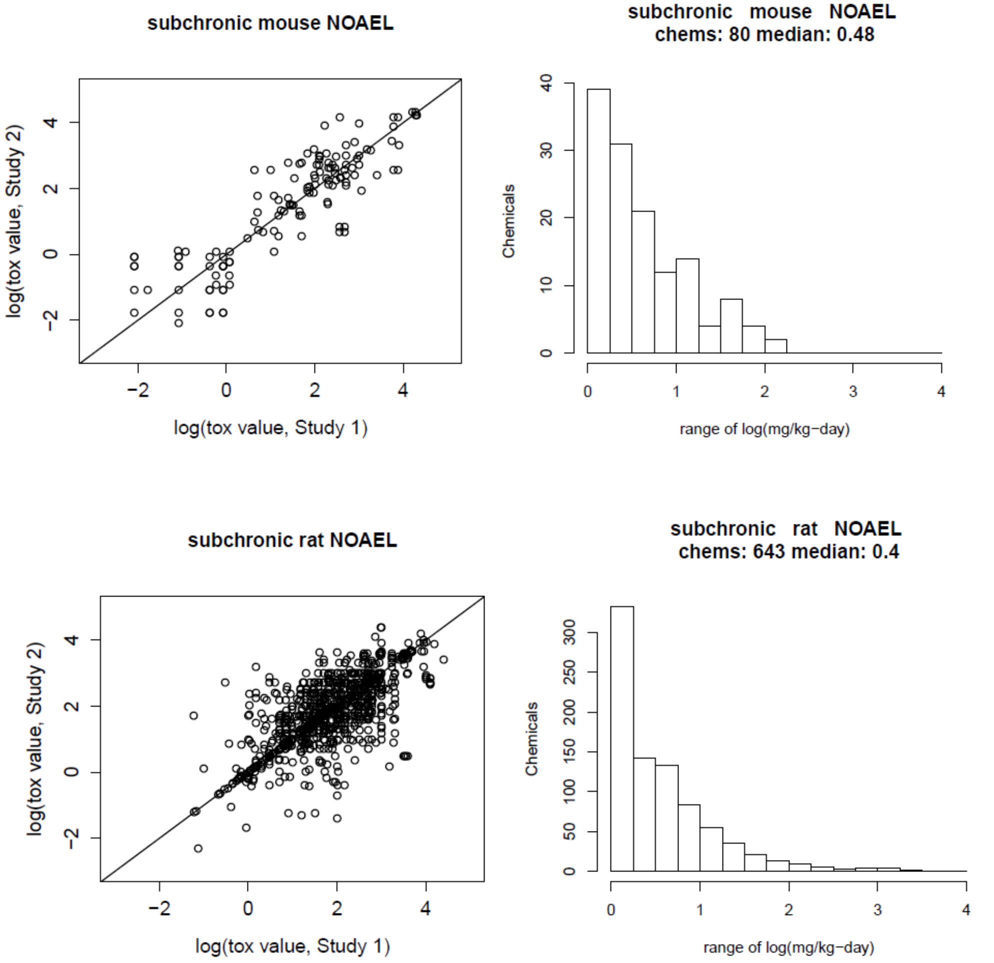
Overall Statistics

source	chemicals	LEL	NEL	BMD	LDx	RfD	LEC	NEC	LCx	BMC	RfC	Cancer
ECOTOX	6807	1601	2076	0	3041	0	1337	1594	4781	3	0	0
ECHA IUCLID	4943	1908	3896	30	4601	0	1610	3353	4047	27	0	0
ECHA	4716	746	2279	5	2048	0	950	2930	2412	1	0	0
TEST	4410	0	0	0	4410	0	0	0	0	0	0	0
EFSA	2875	148	980	96	475	437	1	399	459	0	0	0
COSMOS	1146	598	633	0	872	0	0	0	0	0	0	0
HPVIS	883	276	556	0	752	0	11	12	298	0	0	0
ToxRefDB	868	864	867	0	0	0	0	0	0	0	0	0
RSL	804	0	0	0	0	646	0	0	0	0	229	236
WHO IPCS	580	0	0	0	580	0	0	0	0	0	0	0
Wignall	574	97	441	83	0	512	0	0	0	36	93	23
HESS	530	480	530	0	0	0	0	0	0	0	0	0
IRIS	444	66	254	40	2	366	0	0	0	23	89	71
Pennsylvania DEP	394	0	0	0	0	326	0	0	0	0	126	144
NIOSH	390	0	0	0	0	0	0	0	0	0	0	0
Cal OEHHA	389	0	0	0	0	0	0	0	0	0	0	276
EPA OPP	389	0	0	0	0	387	0	0	0	0	0	46
PPRTV (ORNL)	305	77	163	75	1	271	1	0	0	33	86	45
EPA AEGL	269	0	0	0	0	0	0	0	0	0	0	0
HEAST	212	38	172	0	0	197	0	0	0	1	33	0
ATSDR	196	0	196	0	0	0	0	0	0	0	0	0
OW Drinking Water Standards	194	0	0	0	0	179	0	0	0	0	0	0
PPRTV (NCEA)	176	69	101	55	0	160	0	0	0	2	68	0
DOE ECORISK	156	156	156	0	0	0	0	0	0	0	0	0
Alaska DEC	144	0	0	0	0	124	0	0	0	0	49	69
DOE Wildlife Benchmarks	96	74	96	0	0	0	0	0	0	0	0	0
Health Canada	60	0	0	0	0	0	0	0	0	0	0	14
HAWC	36	34	30	0	0	0	0	0	0	0	0	0

Using ToxValDB to understand Uncertainty / Variability of *In Vivo* data

- Experimental variability
  - Species, strain, dose range, dose spacing
- Experimental error
- Statistical power issues
  - Too few animals to see weak or rare effect
- Reporting bias
  - Was an effect negative or not looked for?
- Observer bias
  - Less severe phenotypes not reported when more severe ones are present
- Diagnostic terminology drift
- Data assimilation and analysis
  - Typos, incomplete transcription

Start with uncertainty in in vivo dose metrics



Source of Data: ToxValDB, combines

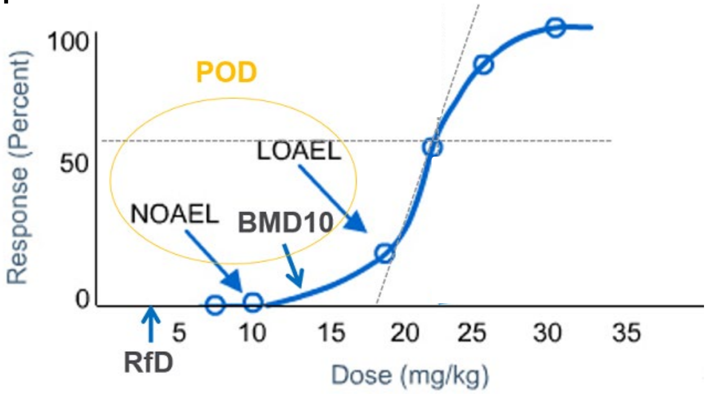
- ECHA
- EFSA
- COSMOS
- IRIS
- PPRTV
- ToxRefDB
- Etc.

Each point is one chemical, one study protocol, one species, one POD type run in two labs

Many instances of PODs differing by 1-2 orders of magnitude

Mammalian POD QSAR Model

Point-of-departure (POD) is the point on the dose-response curve that marks the beginning of a low-dose extrapolation



Goal: To develop quantitative structure-activity relationship (QSAR) models for predicting systemic toxicity PODs incorporating variability in underlying data to derive uncertainty in model predictions

Motivation: Development of faster and efficient alternative (non-animal) methods for risk assessment and screening of a large number of data-poor chemicals

Image from:
http://www.chemsafetypro.com/Topics/CRA/What_is_Point_of_Departure_(POD)_in_Toxicology_and_How_to_Use_It_to_Calculate_Reference_Dose_RfD.html

Prachi Pradeep

DATASET

ToxValDB, a compilation of information on ~4000 unique chemicals from a variety of public data sources including:

- ToxRefDB
- IRIS
- PPRTV (ORNL)
- PPRTV (NCEA)
- ECHA
- COSMOS
- CalEPA
- EPA
- ..and more.

Effect level types:

- LEL, LEC
- LOEL, LOEC
- LOAEL, LOAEC
- NEL
- NOEL, NOEC
- NOAEL, NOAEC
- BMD, BMC, BMC10
- BMDL, BMDL-01, BMDL-05, BMDL-10, BMDL-1SD, BMCL, 'BMCL-5', 'BMCL-10', 'BMCL-1SD’
- PODs

Study Type	Species	Total number of POD values (studies)	Number of unique chemicals
Chronic (CHR)	Rat	13423	3047
	Mouse	4130	690
	Rabbit	342	240
	Rat, Mouse, Rabbit	17895	3221
Subchronic (SUB)	Rat	6696	988
	Mouse	2418	308
	Rat, Mouse	9114	1030
Reproductive (REP)	Rat	2915	425
	Mouse	244	62
	Rat, Mouse	3159	460
Developmental (DEV)	Rat	2472	416
	Rabbit	1540	273
	Rat, Rabbit	4012	511
Subacute (SAC)	Rat	1133	155
ALL (CHR, SUB, REP, DEV, SAC)	**All (Rat, Mouse, Rabbit)**	**36013**	**3762**

Prachi Pradeep

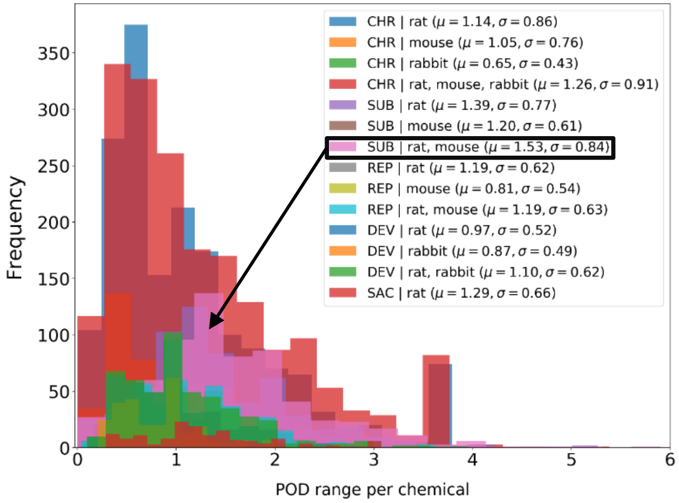
Mammalian POD QSAR Model: MODELING CHALLENGES

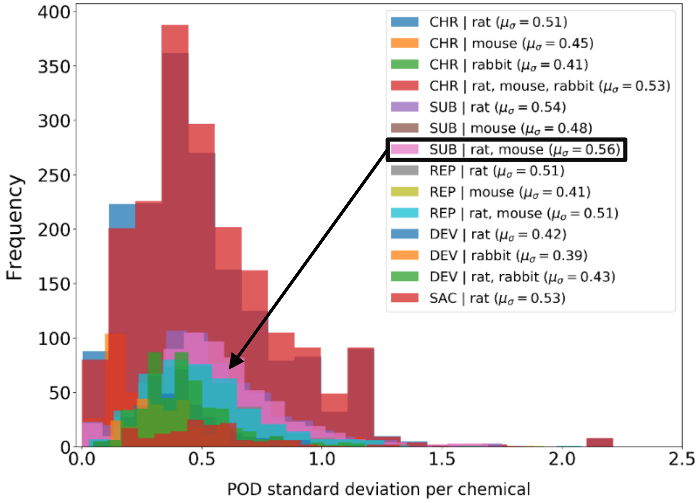
1. Experimental Variability

- Data from different labs (sources) running the "same" experiment may get different answers
- Sources of variability: biological (e.g., test species, environmental conditions) and/or technical (e.g., measurement errors, different experimental protocols)

2. Model Uncertainty

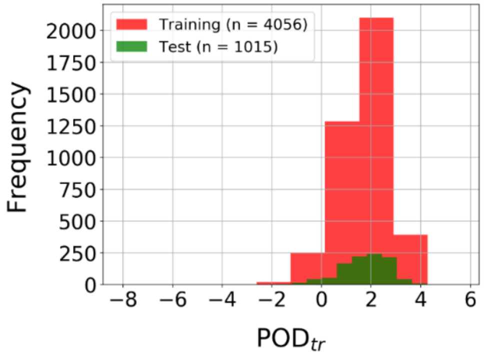
- A model gives a result (a POD), but this is an estimate of the "true" POD. The true POD is mostly unknown.
- Uncertainty in the evaluation data will lead to uncertainty in the model and our estimate of its quality

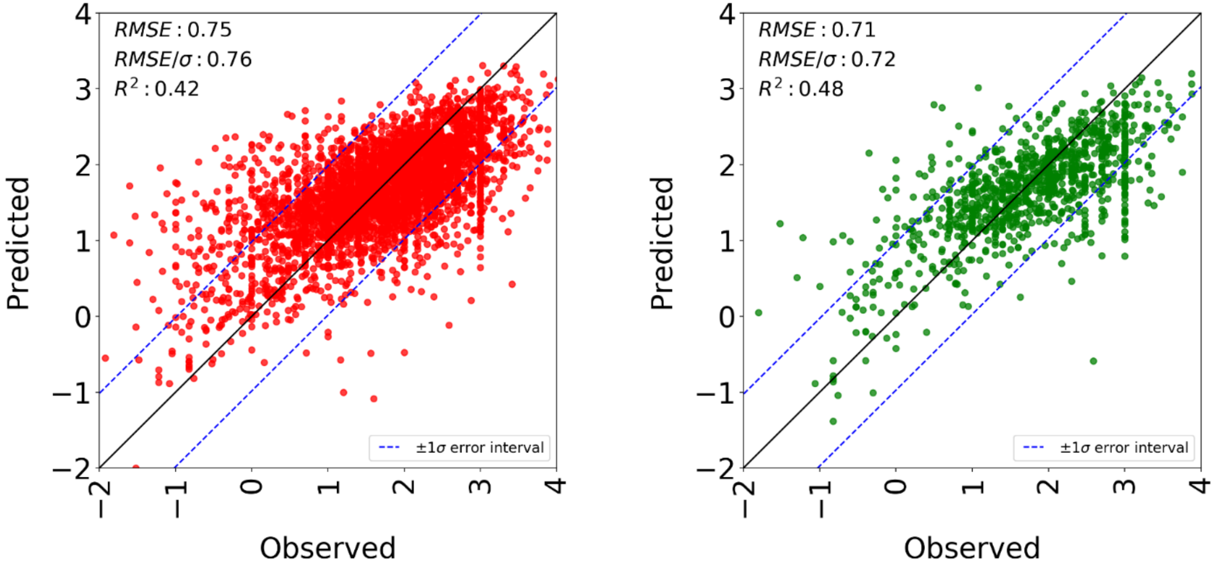




Prachi Pradeep

Mammalian POD QSAR Model: Point-estimate Models



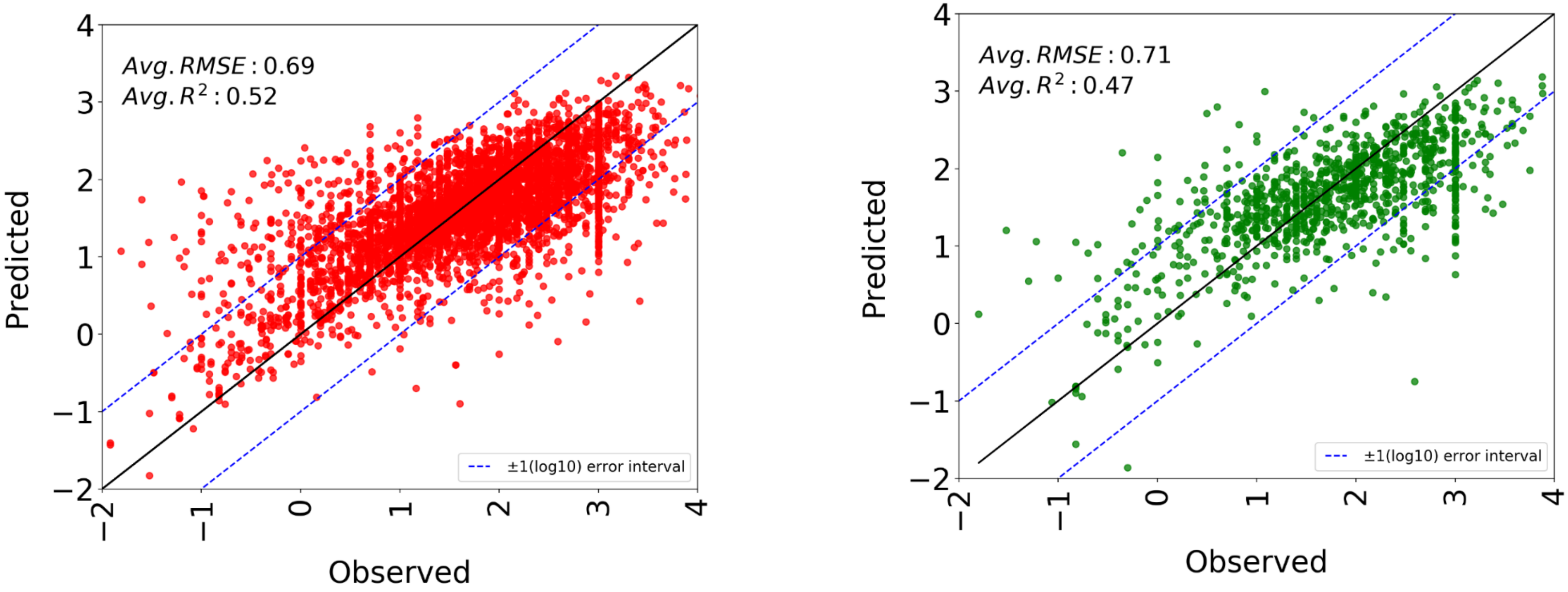


Observed versus predicted POD values (Log$_{10}$-transformed)

5-fold internal cross-validation (red scatter plot) and external validation (green scatter plot) for the best model (random forest) developed using a combination of all study types (CHR, SUB, DEV, REP and SUB) and all species (rat, mouse and rabbit) and using species and study type as additional descriptors in the model.

Prachi Pradeep

Mammalian POD QSAR Model: Point-estimate with Confidence Intervals Models



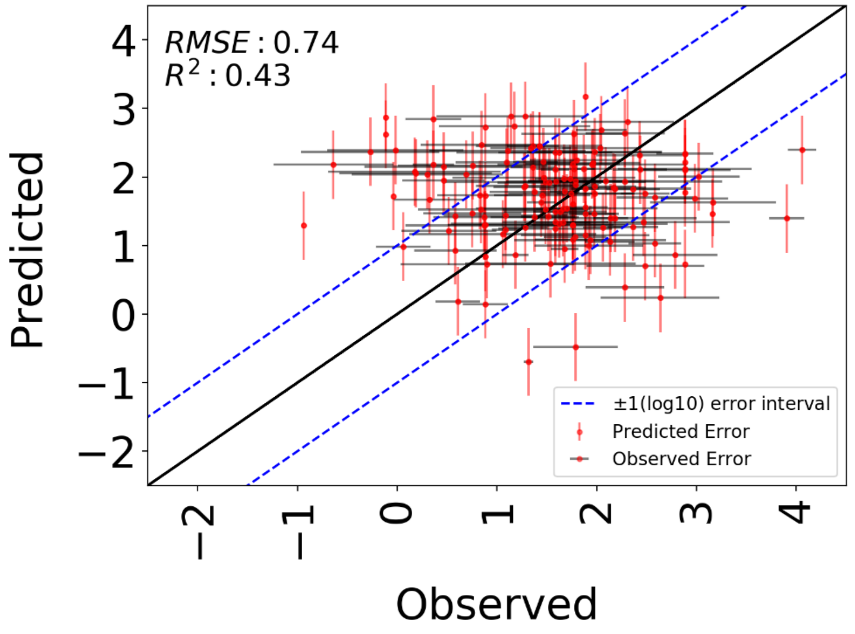
Observed versus predicted POD values (Log$_{10}$-transformed)

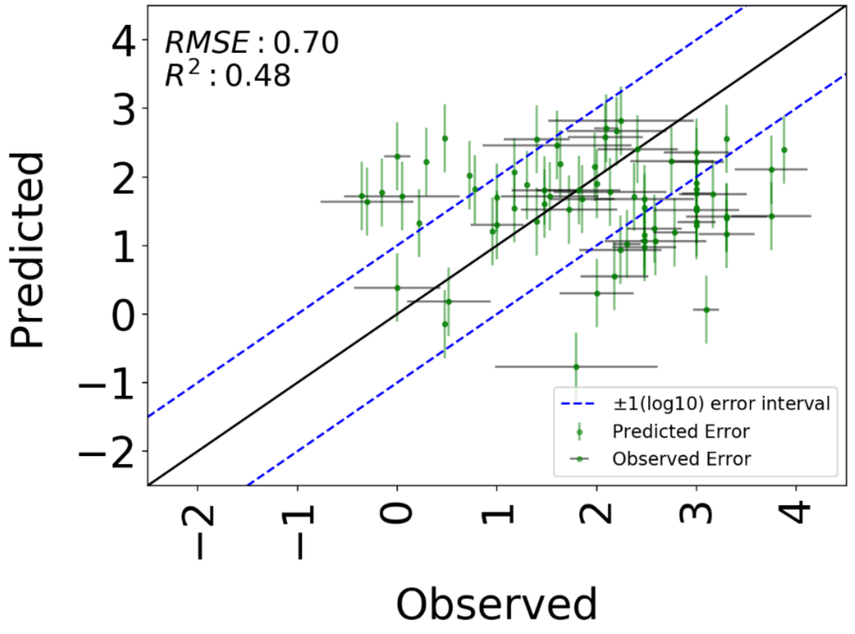
Ref.
Training
RMSE:
0.75
R2: 0.42

Ref. Test
RMSE:
0.71
R2: 0.48

Prachi Pradeep

Mammalian POD QSAR Model: Point-estimate with Confidence Intervals Models





Ref. Training RMSE: 0.75 R2: 0.42

Observed versus predicted POD values (Log$_{10}$-transformed)

50 chemicals were selected randomly and plotted for each dataset.

Ref. Test RMSE: 0.71 R2: 0.48

Prachi Pradeep

Mammalian POD QSAR Model: SUMMARY

1. **Point-estimate model** results demonstrate that independent study type and species combinations did not result in significantly better models than combining the data for all the classes and species together.
   - The RMSE for the all the models are within the variance in the underlying POD data.
   - Enrichment analysis results demonstrate the utility of these models for chemical screening and prioritization efforts.
2. **Point-estimate with balanced dataset models** results show improvement in the training set results but did not show improved results on the external test sets.
3. **Point-estimate with confidence interval models** presented a technique to estimate uncertainty associated with model predictions. The results demonstrate the impact of variability in training data (experimental POD) on uncertainty associated with model results.

Prachi Pradeep

Fish Toxicity QSAR model

- QSAR model for points of departure in fish (multi-species)

- Use all available ToxvalDB data where possible

- Use study covariates as features

- Two models:
  - Acute LC_{50} (“LC_{50} Model”)
  - Any duration NOEC/LOEC/LC_0/MATC Growth/Mortality/Reproductive (“NOEC Model”)

Thomas Sheffield

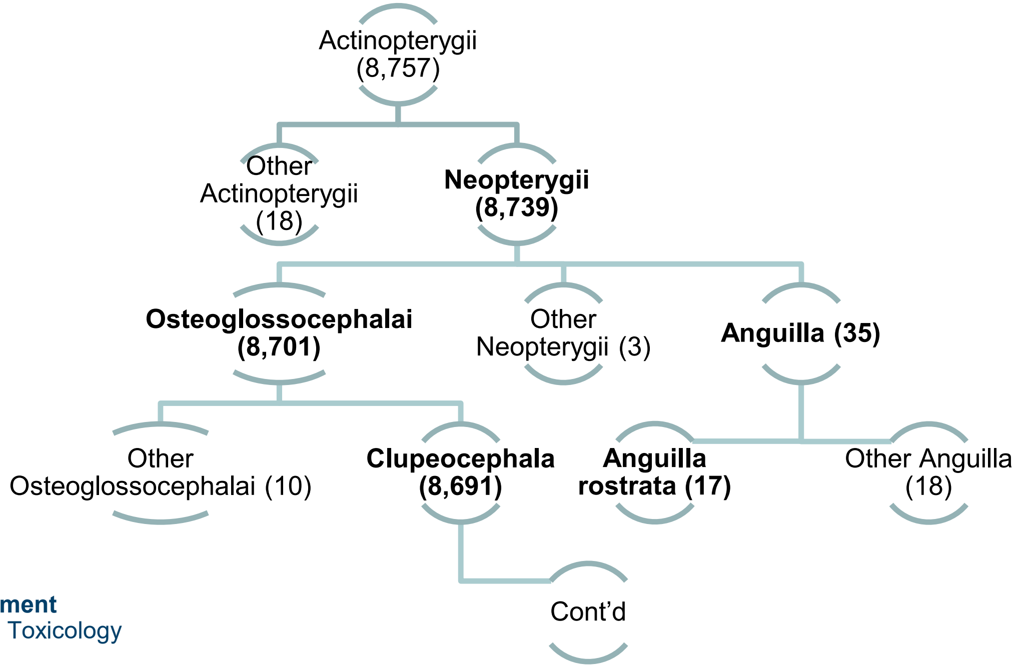
Fish Toxicity QSAR model: Data

- Drawn from ECOTOX (89%) and ECHA (11%) databases
- Substantial cleaning required
  - Standardize study covariates
    - Species, endpoint type, study type, study duration class, exposure route, endpoint units
  - Drop rare, incongruous, or suspect experiment types
  - Merge salts and stereoisomers
- Final LC_{50} model: 34,645 experiments, 2,656 chemicals, and 358 species
- Final NOEC model: 14,484 experiments, 1,926 chemicals, and 221 species

Thomas Sheffield

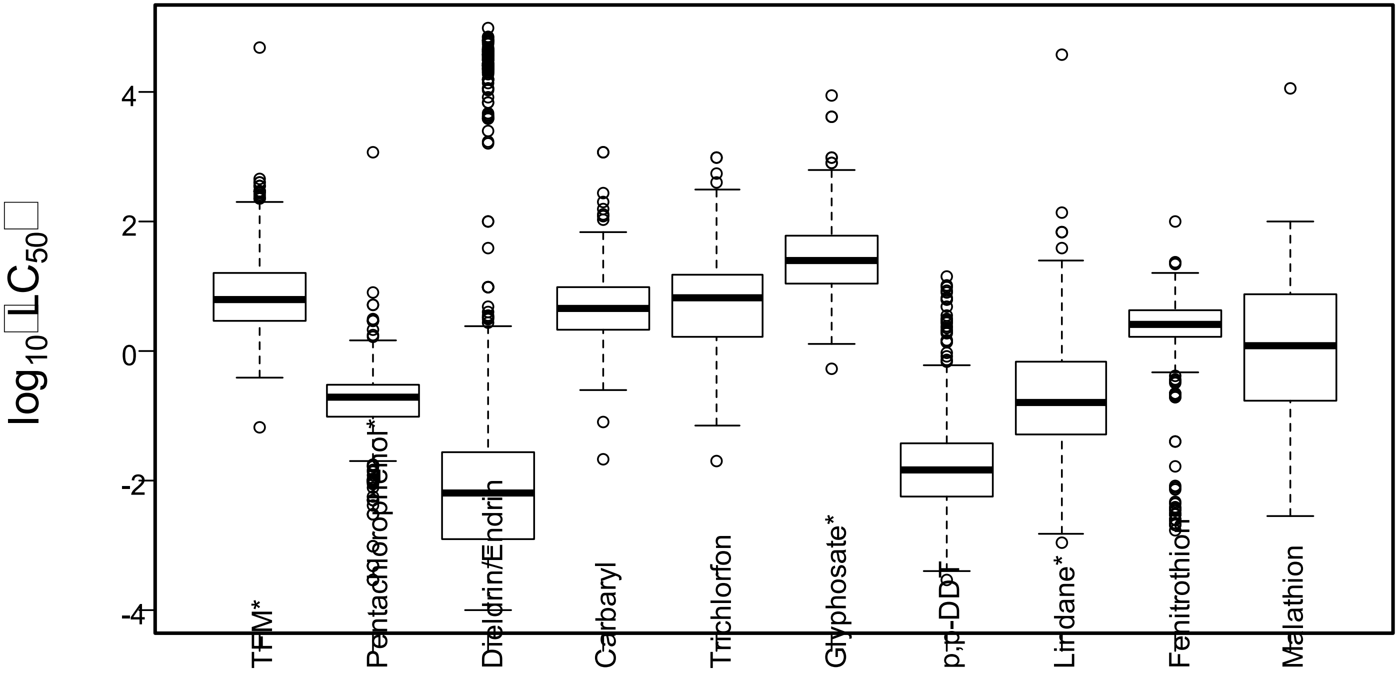
Fish Toxicity QSAR model: Features

- OPERA Physiochemical Properties (11)
- PaDEL Descriptors (1,444)
- Experimental Covariates (~1,300)
  - Exposure route and taxonomy groups (LC_{50} and NOEC)
  - Study type, endpoint type, duration class (NOEC only)


Actinopterygii (8,757)
Other Actinopterygii (18)
Neopterygii (8,739)
Osteoglossocephalai (8,701)
Other Neopterygii (3)
Anguilla (35)
Other Osteoglossocephalai (10)
Clupeocephala (8,691)
Anguilla rostrata (17)
Other Anguilla (18)
Cont'd

Thomas Sheffield

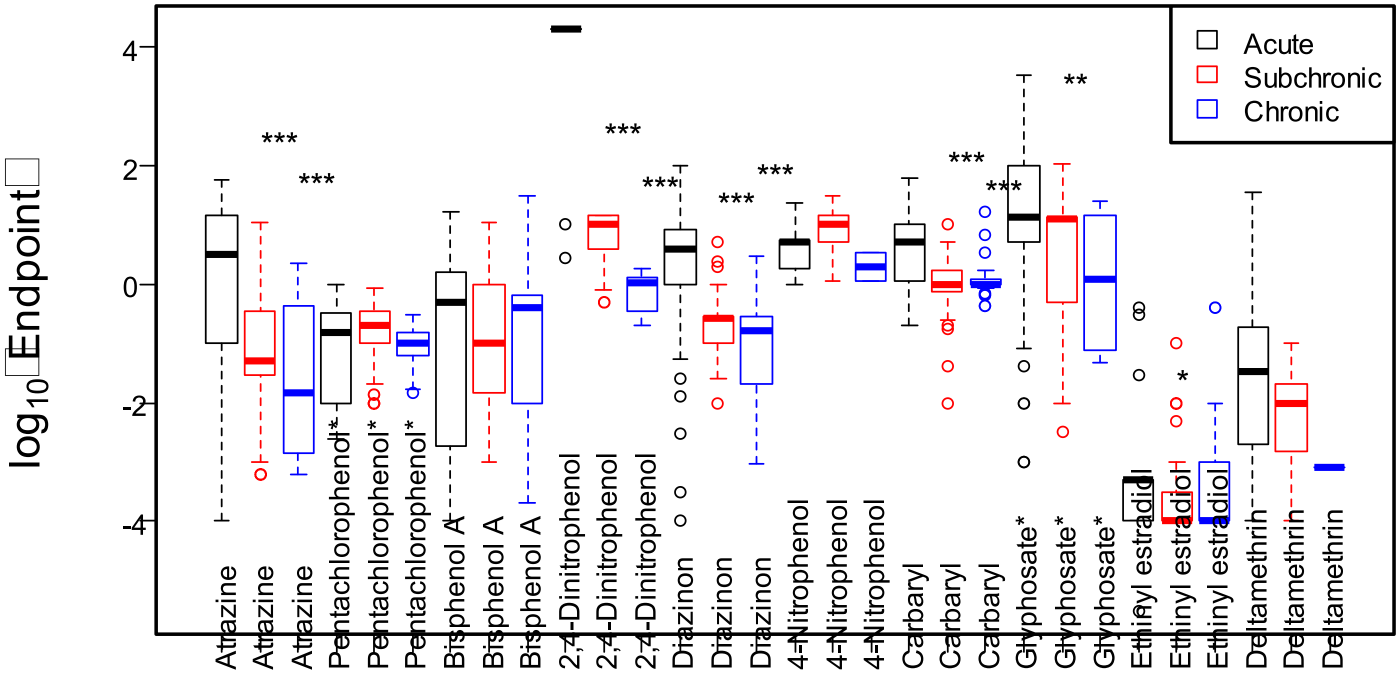
Fish Toxicity QSAR model: LC_{50} Most Common Chemicals



* denotes merged salts/stereoisomers

Thomas Sheffield

Fish Toxicity QSAR model: NOEC Most Common Chemicals



Thomas Sheffield

Fish Toxicity QSAR model: Data Variability

- Average standard deviation of chemicals with ten or more entries:
  - LC$_{50}$ model: 0.53 $\log_{10}$(mg/L) (468 chemicals)
  - NOEC model: 0.78 $\log_{10}$(mg/L) (319 chemicals)

- Average standard deviation of experiment groups (same study covariates & chemical) with ten or more entries:
  - LC$_{50}$ model: 0.41 $\log_{10}$(mg/L) (638 exp. groups)
  - NOEC model: 0.35 $\log_{10}$(mg/L) (105 exp. groups)

Thomas Sheffield

Fish Toxicity QSAR model: LC_{50} Performance Summary

- No apparent overfitting
- Full model and fast model perform the same
- A little more error when predicting experiment groups vs. chemical average
- Overall, RMSE ~ 0.8 and R^2 ~ 0.6
- About 81% of chemicals predicted within one order of magnitude

Thomas Sheffield

Fish Toxicity QSAR model: NOEC Performance Summary

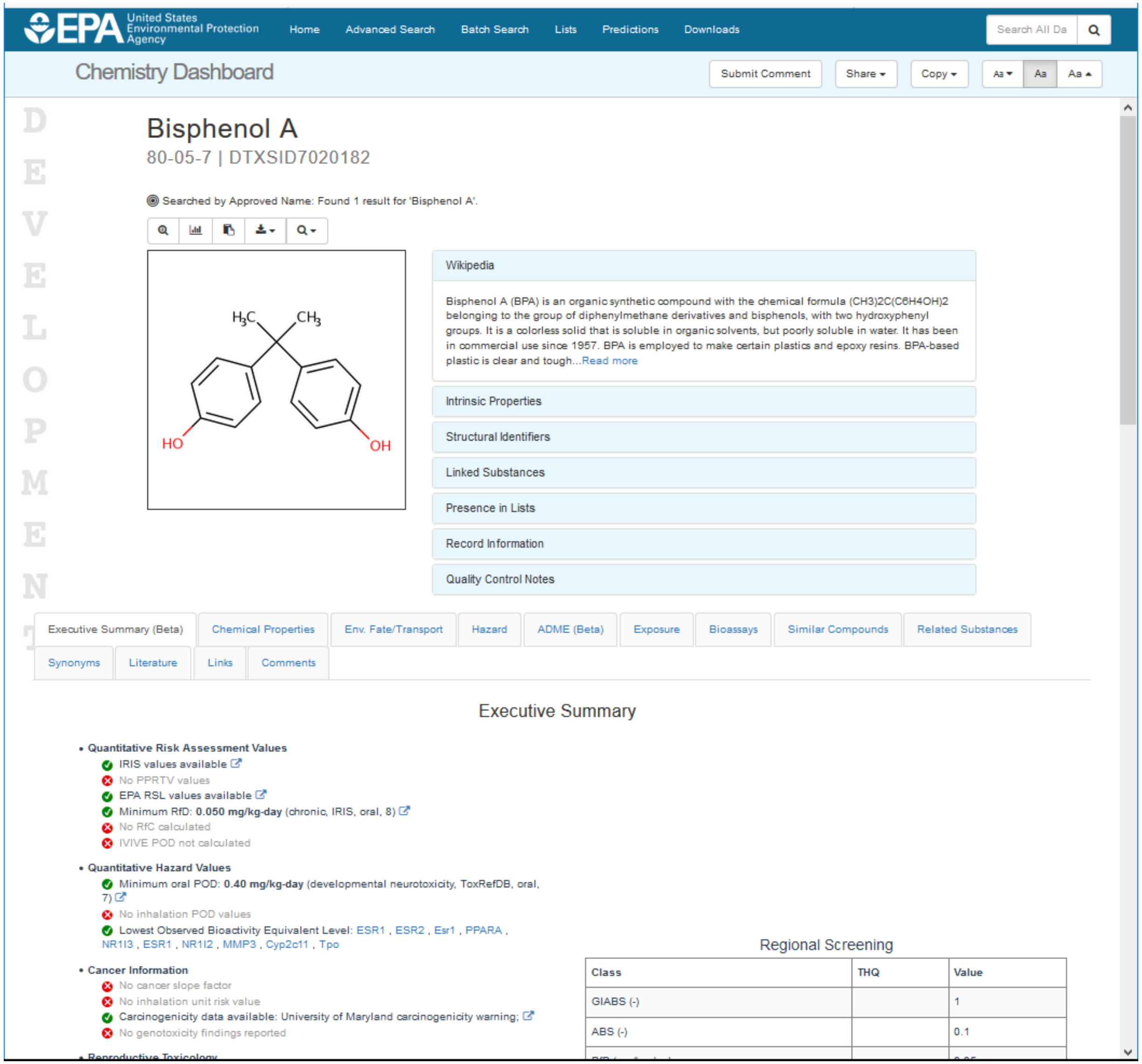
- Similar behavior to LC_{50}
- Chronic study performance similar to overall performance
- Overall, RMSE ~ 1.0 and R^2 ~ 0.6
- About 76% of chemicals predicted within one order of magnitude

Thomas Sheffield

Ongoing work

- Further automated data cleaning
  - E.g. matching study type with study duration
- Developing a manual QC process
- Continue to bring in new data sources
  - Working with ECHA to access all REACH data
- Redesign of Comptox Chemicals Dashboard view of ToxValDB
- Accessing literature data – a big challenge
- Multiple ongoing applications
  - QSAR models
  - Prioritization projects
  - RapidTox

Comptox 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 (CH3)2C(C6H4OH)2 belonging to the group of diphenylmethane derivatives and bisphenols, with two hydroxyphenyl 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

Linked Substances

Presence in Lists

Record Information

Quality Control Notes

Executive Summary

- **Quantitative Risk Assessment Values**
  - ✓ IRIS values available
  - ✗ No PPRTV values
  - ✓ EPA RSL values available
  - ✓ Minimum RfD: **0.050 mg/kg-day** (chronic, IRIS, oral, 8)
  - ✗ No RfC calculated
  - ✗ IVIVE POD not calculated
- **Quantitative Hazard Values**
  - ✓ Minimum oral POD: **0.40 mg/kg-day** (developmental neurotoxicity, ToxRefDB, oral, 7)
  - ✗ No inhalation POD values
  - ✓ Lowest Observed Bioactivity Equivalent Level: ESR1 , ESR2 , Esr1 , PPARA , NR1I3 , ESR1 , NR1I2 , MMP3 , Cyp2c11 , Tpo
- **Cancer Information**
  - ✗ No cancer slope factor
  - ✗ No inhalation unit risk value
  - ✓ Carcinogenicity data available: University of Maryland carcinogenicity warning;
  - ✗ No genotoxicity findings reported
- **Reproductive Toxicology**

Regional Screening

Class	THQ	Value
GIABS (-)		1
ABS (-)		0.1

Designed for OLEM application

Quick view of available data

Access to Data

- Comptox Chemicals Dashboard
  - URL: https://comptox.epa.gov
- FTP data download
  - Currently internal version only
- Contact information
  - Richard Judson
  - Judson.richard@epa.gov
  - 919-541-3085

National Center for Computational Toxicology

Collaborators on the ToxValDB Project

Duncan McPherson

Katie Paul Friedman

Sean Watford

Tony Williams

Tommy Cathey

Jeff Edwards

Chris Grulke

Doris Smith

Jamie Vail

Amar Singh

Grace Patlewicz

Prachi Pradeep

Thomas Sheffield

Nathan Rush