EPA
United States
Environmental Protection
Agency

Consensus Modeling in Support of a Semi-Automated Read-Across Application

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

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
COMPUTATIONAL TOXICOLOGY

SOT 2016, New Orleans

March 17, 2016

Problem Statement

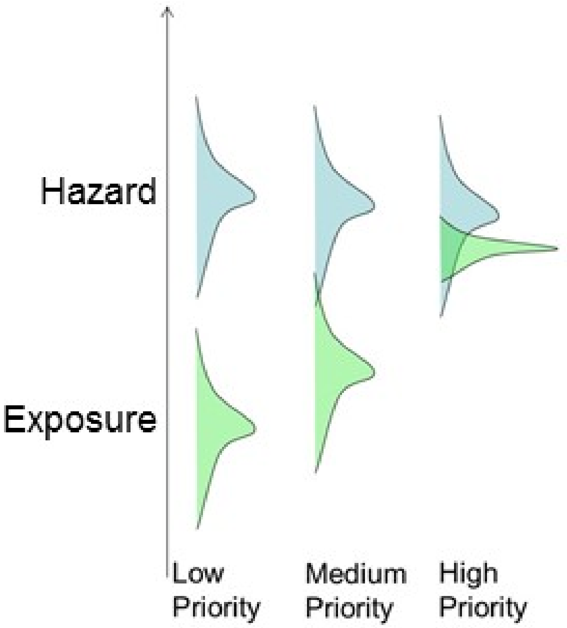
- Too many chemicals, too little data
  - There are tens of thousands of man-made chemicals in the environment, and few of them are thoroughly tested for potential toxicity
- Need to use data-gap-filling methods / models
  - Read-across, QSAR, QBAR, systems models
- All data and models are subject to errors, uncertainty and noise
- Need to develop methods to manage these issues

Key Points

- Goal is to build predictive models in the presence of noisy data

- Recognize and quantify uncertainty
- Build models on the best (most reproducible) data
- Combine multiple imperfect models together (consensus)
- Build local models where possible

Never despair:
You may not know much but you never know nothing



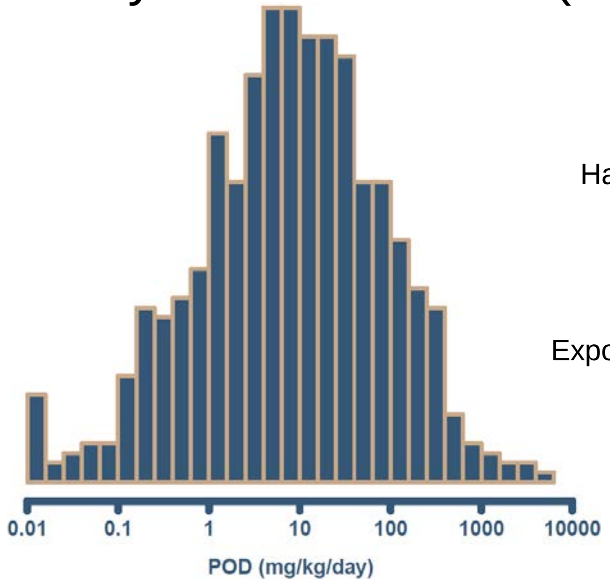
What are the limits on predictivity?

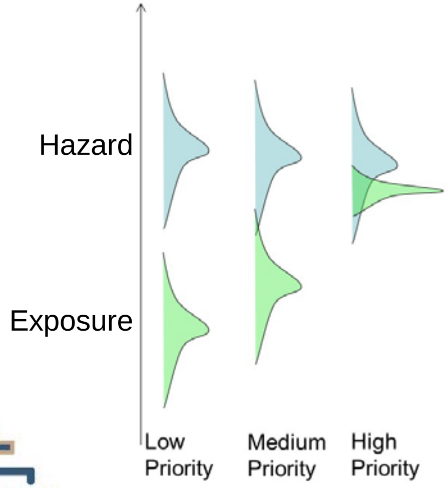
Example using predictions of PODs

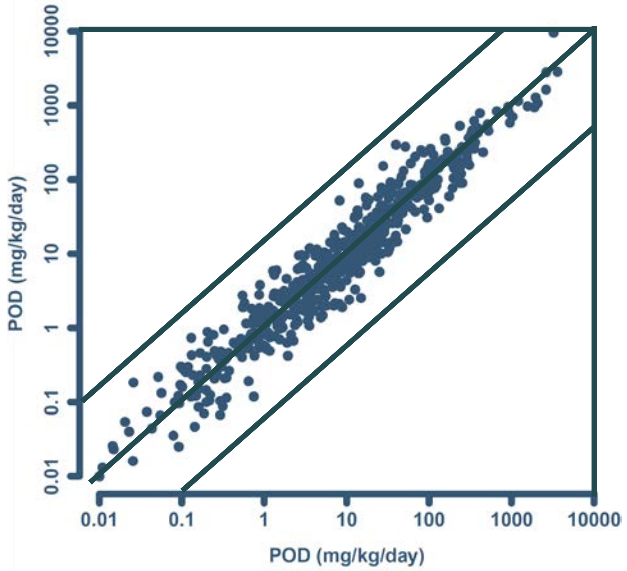
Quantify uncertainty

Worst case: Predict the mean of all chemicals – at worst a prediction will be off by a factor of 1000 (3 logs)

Best case: RMSE cannot be better on average than 0.3 log units due to typical wide dose spacing







Martin et al. (in prep)

Case Study: Endocrine Disruption

- EPA has to test ~10,000 chemicals in the EDSP
  - Tier 1 battery can run at ~50/year at $1M/chemical
  - 100+ years, billions of dollars
  - Even the tier 1 guideline studies are <u>imperfect</u>

Quantify uncertainty

Consensus Model

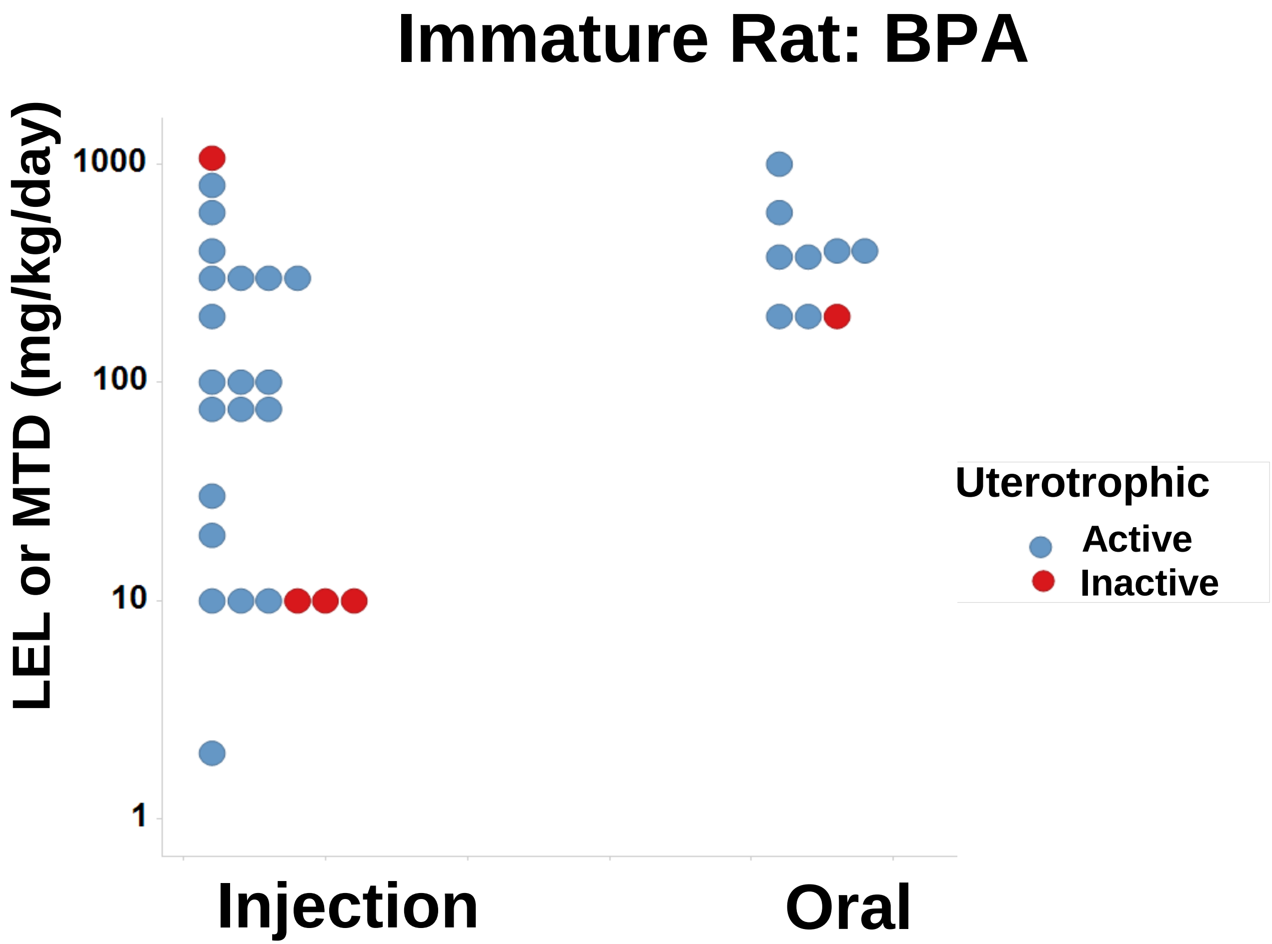
- Proposed approach is to use a <u>combination</u> of methods
  - Tier 1, in vivo read-across, HTS, in vitro-based models, QSAR
  - <u>Combine</u> staged replacement of tests with prioritization
  - But the new approaches are also <u>imperfect</u>

- Today focus on estrogen receptor activity

In vivo guideline study uncertainty

26% of chemicals tested multiple times in the uterotrophic assay gave discrepant results

Quantify uncertainty



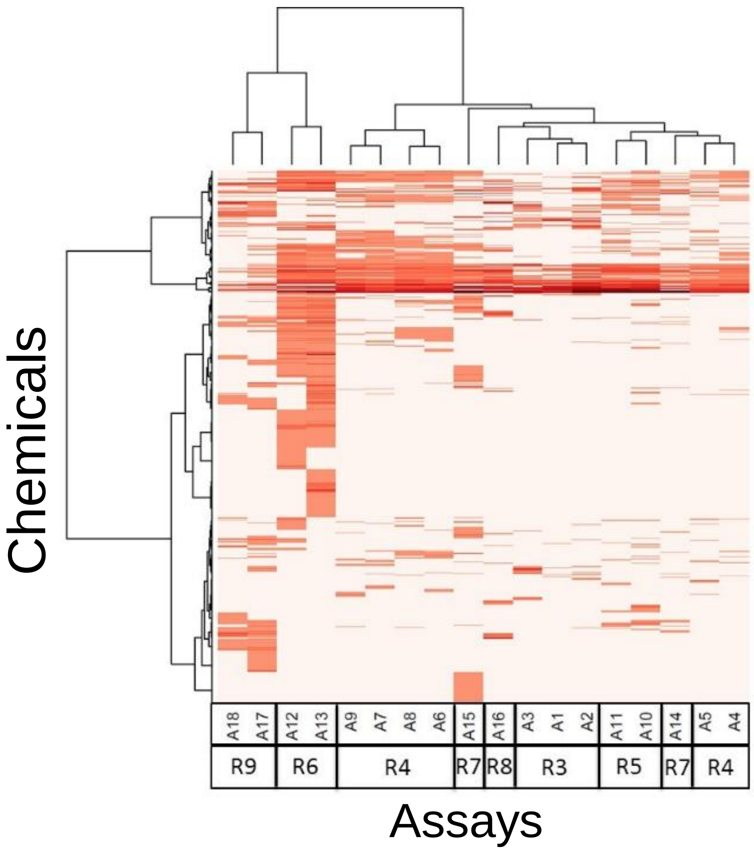
Kleinstreuer et al. EHP 2015

In vitro assays also have false positives and negatives

Quantify uncertainty

Consensus Model

Assays cluster by technology, suggesting technology-specific non-ER bioactivity



Much of this “noise” is reproducible

- “assay interference”
- Result of interaction of chemical with complex biology in the assay

EDSP chemical universe is structurally diverse

-Solvents

-Surfactants

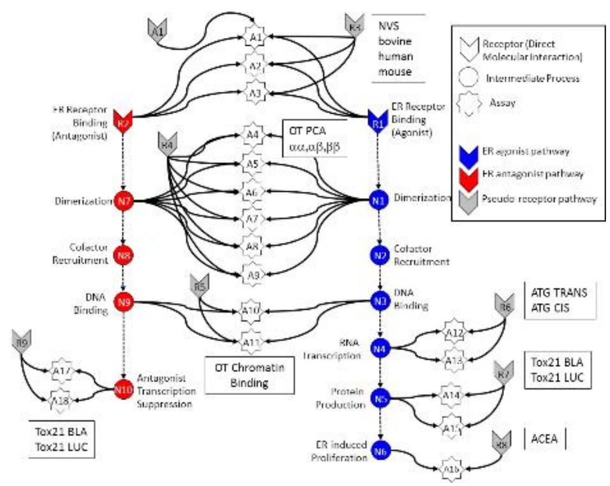
-Intentionally cytotoxic compounds

-Metals

-Inorganics

-Pesticides

-Drugs



Judson et al: ToxSci (2015)

Assay-to-assay variation

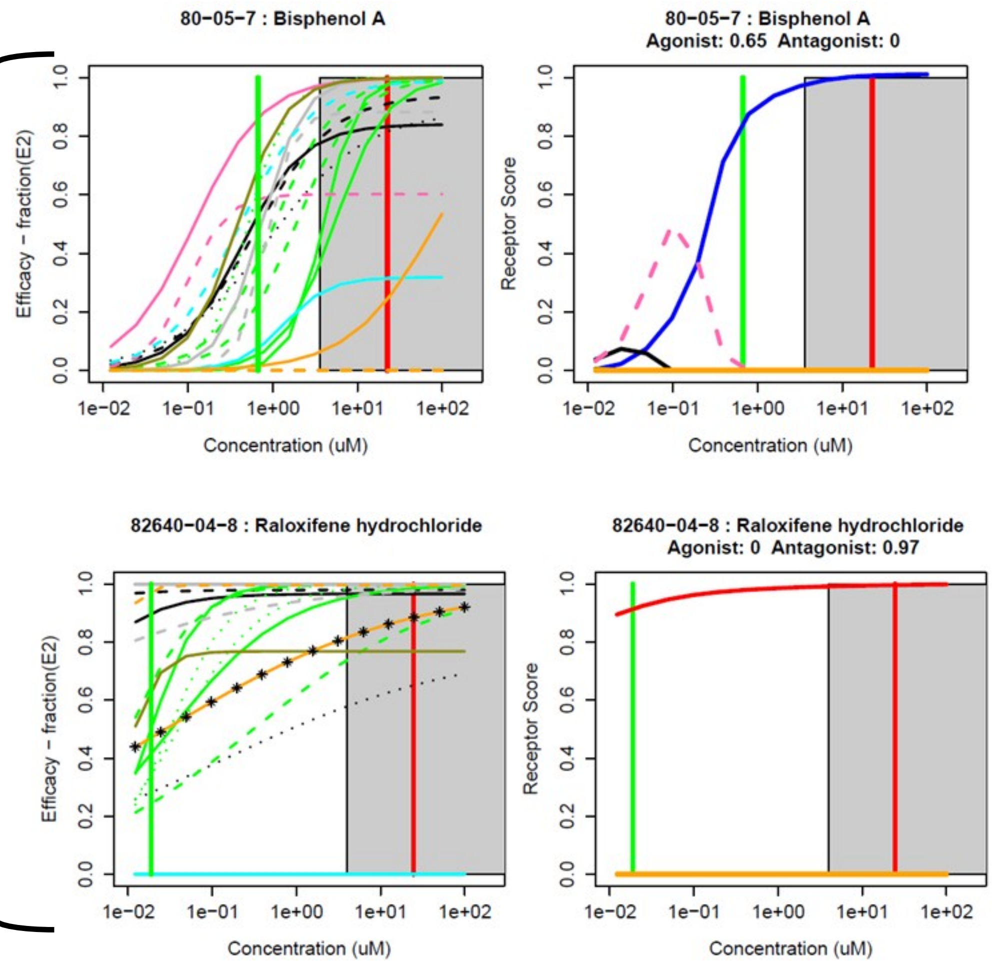
Quantify uncertainty

Consensus Model

All appropriate assays are active but efficacy and potency vary

“Noise” or real variation in biology between cell types?

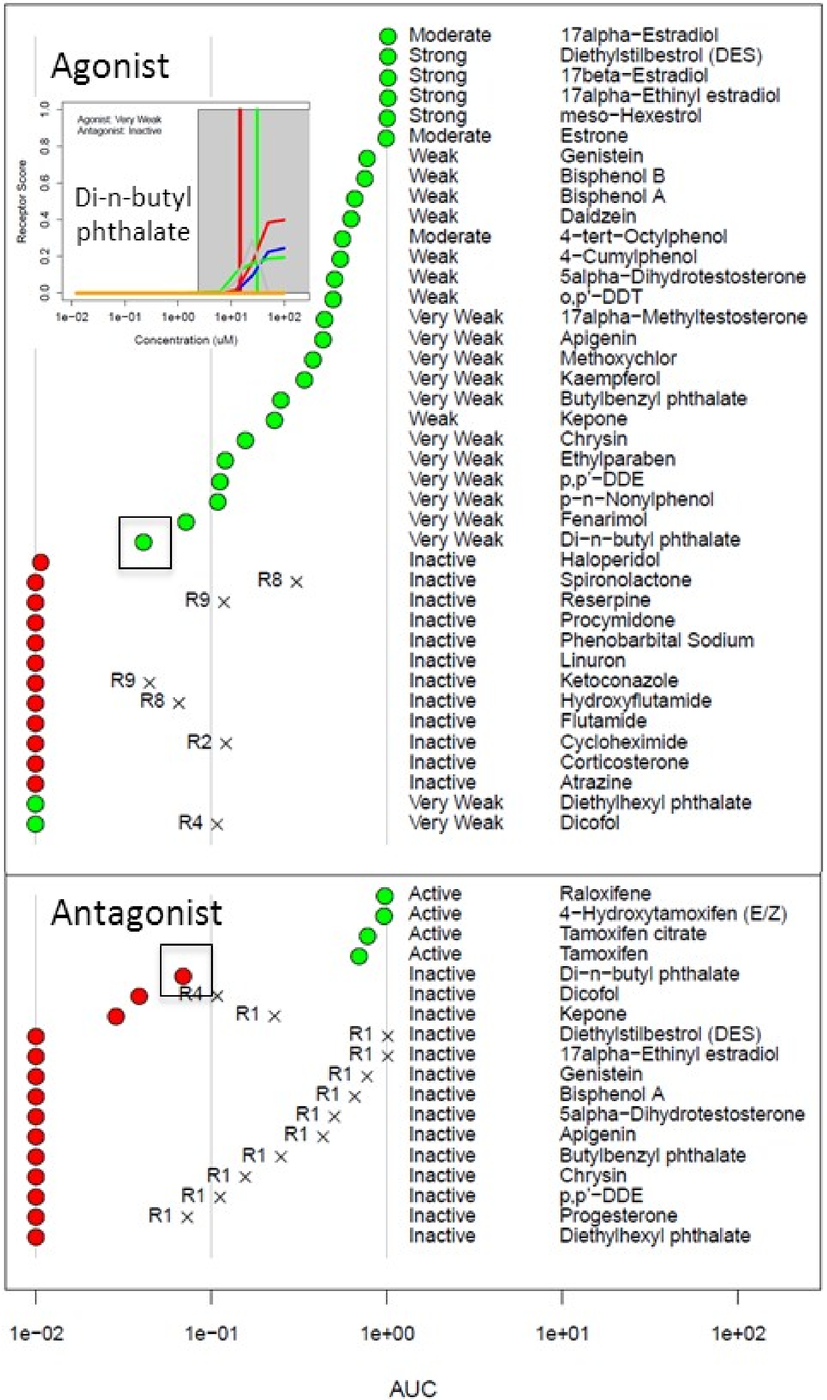
Assay Data | Integrated Model



Agonist

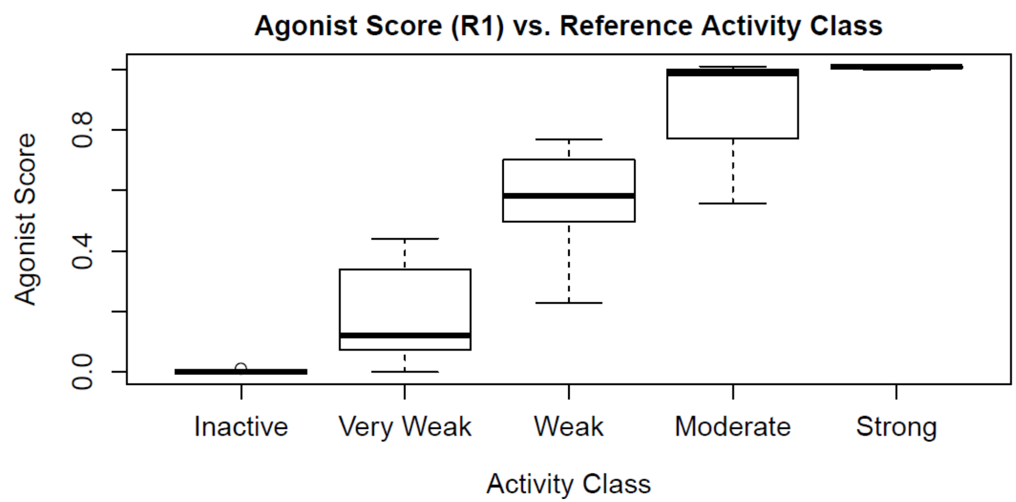
Antagonist

Judson et al: ToxSci (2015)



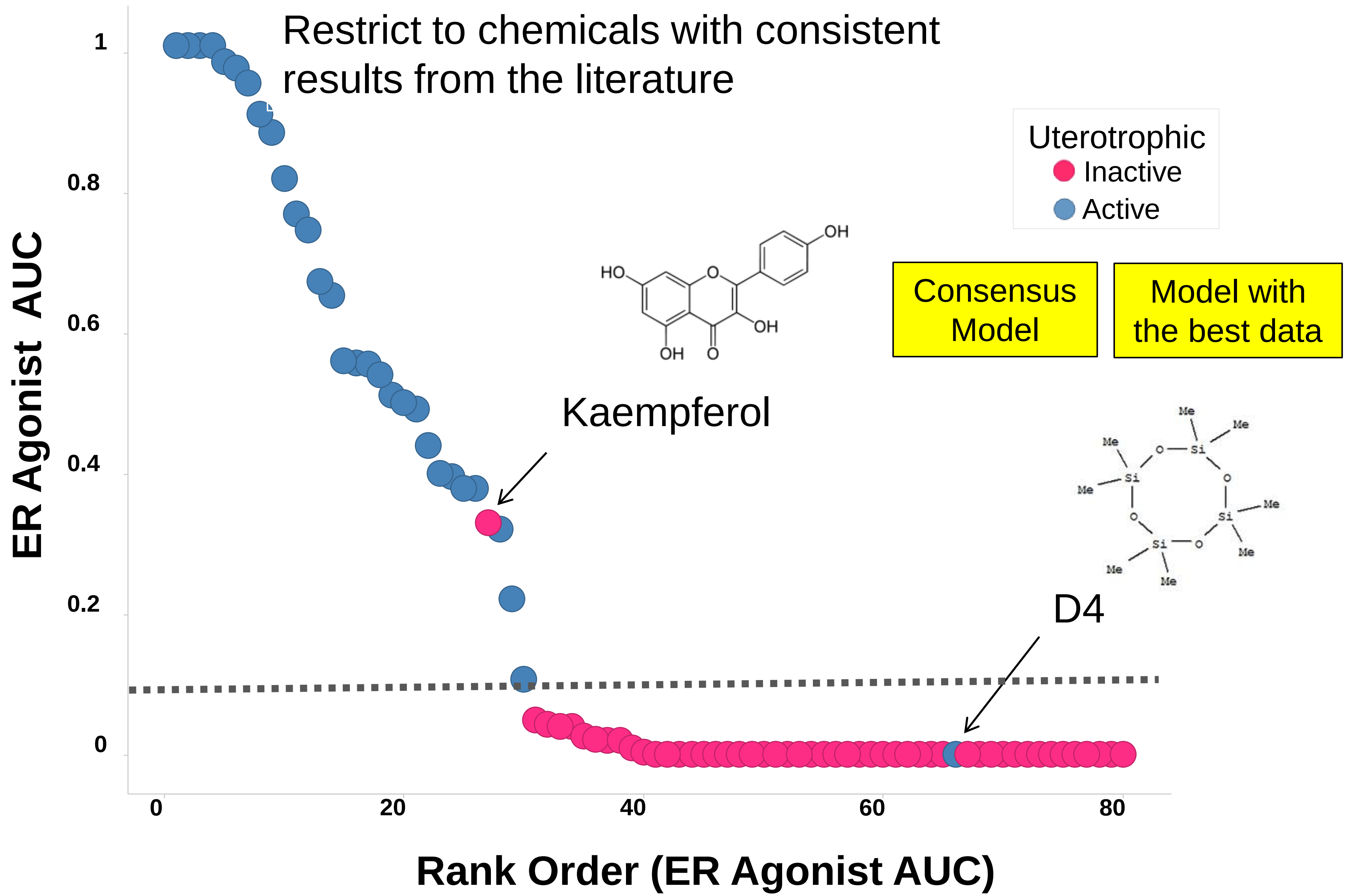
Despite assay-to-assay variation, model or "average performance" predicts reference chemicals accurately

Consensus Model	Model with the best data



Judson et al: ToxSci (2015)

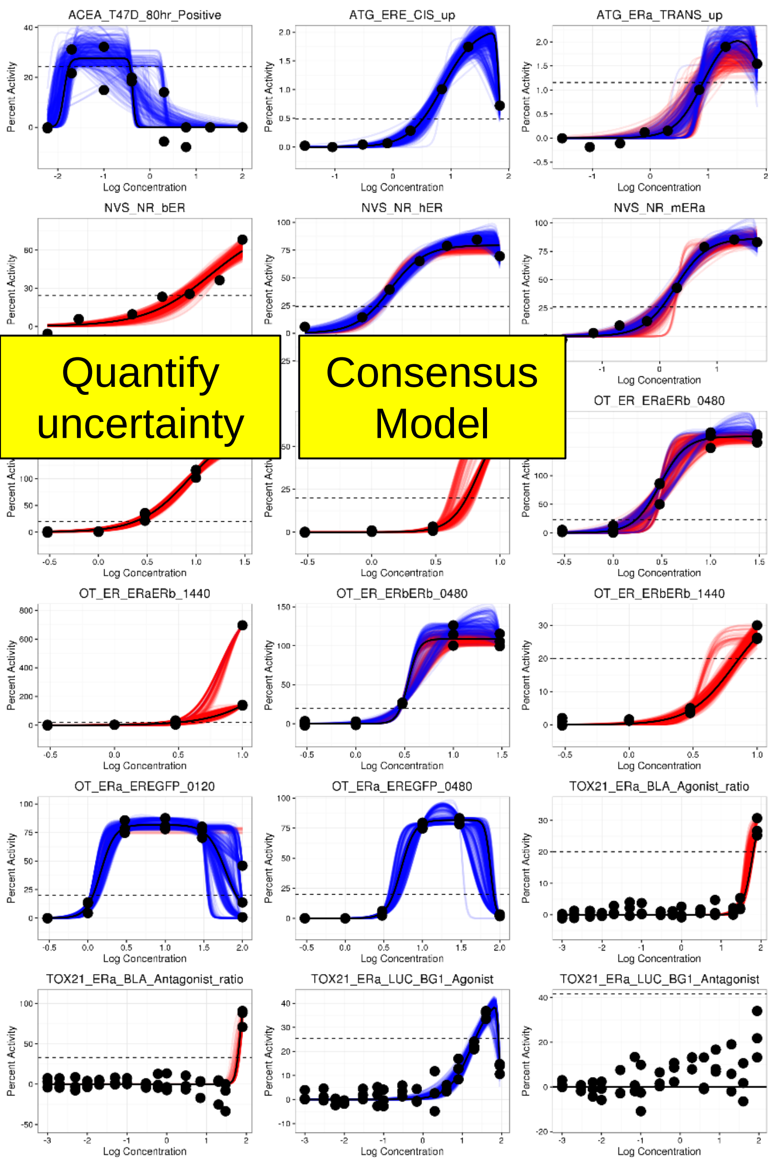
Model also predicts *in vivo* uterotrophic assay as well as uterotrophic predicts uterotrophic



Browne et al. ES&T (2015)

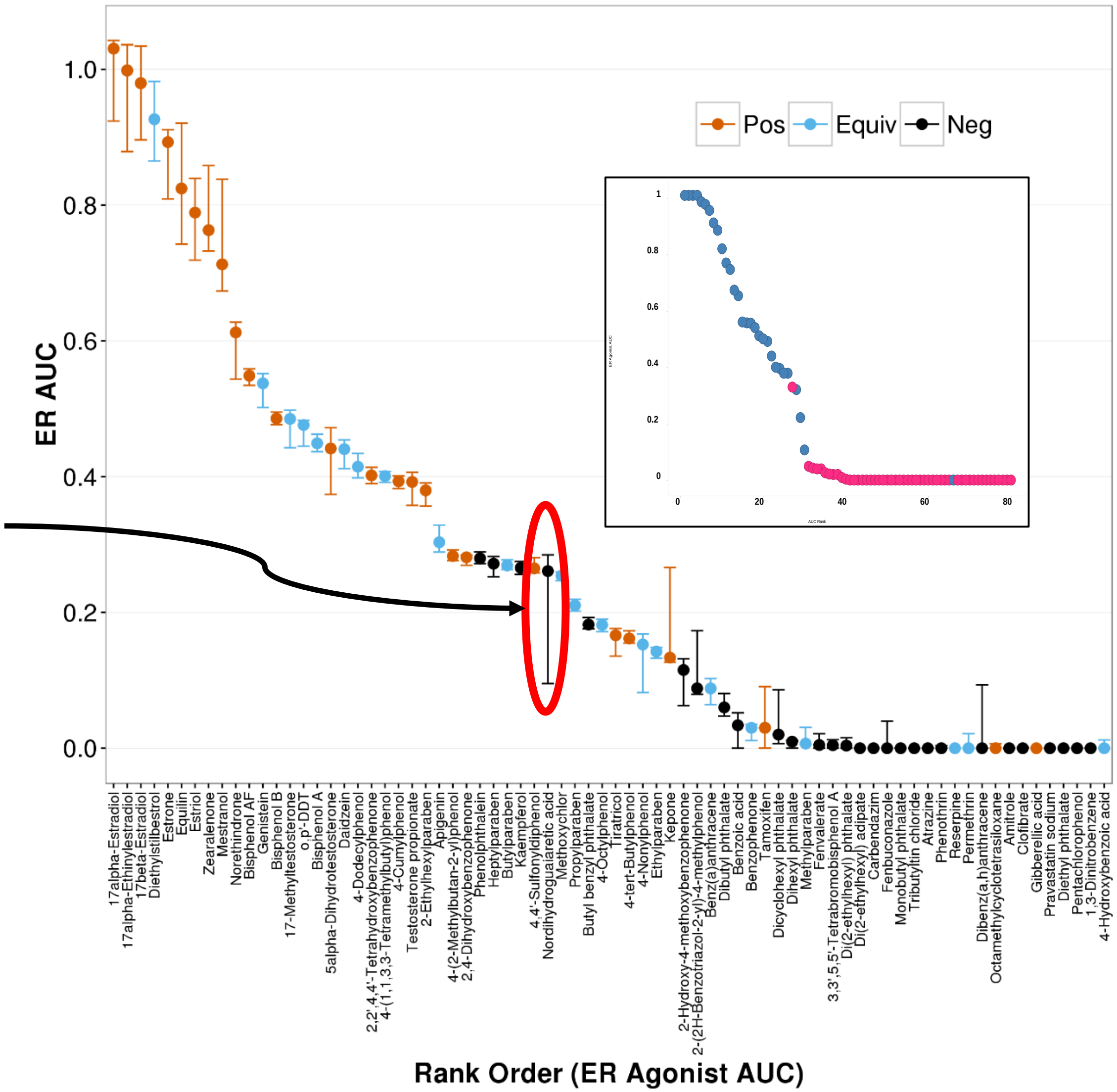
United States Environmental Protection Agency

Explicitly Add Uncertainty to *In Vitro* Assay Data



Quantify uncertainty

Consensus Model



Watt et al. (in prep)

CERAPP: using QSAR for further prioritization

- Collaborative Estrogen Receptor Activity Prediction Project
- Goals:
  - Use ToxCast ER score (or other data) to build many QSAR models
  - Use consensus of models to prioritize chemicals for further testing
- Assumptions
  - ToxCast chemicals cover enough of chemical space to be a good "**global**" training set
  - Consensus of many models will be better than any one individually
- Process
  - Curate chemical structures
  - Curate literature data set
  - Build many models
  - Build consensus model
  - Evaluate models and consensus

Consensus Model

Model with the best data

Mansouri, et al. EHP 2016

Consensus of models and data helps QSAR model accuracy

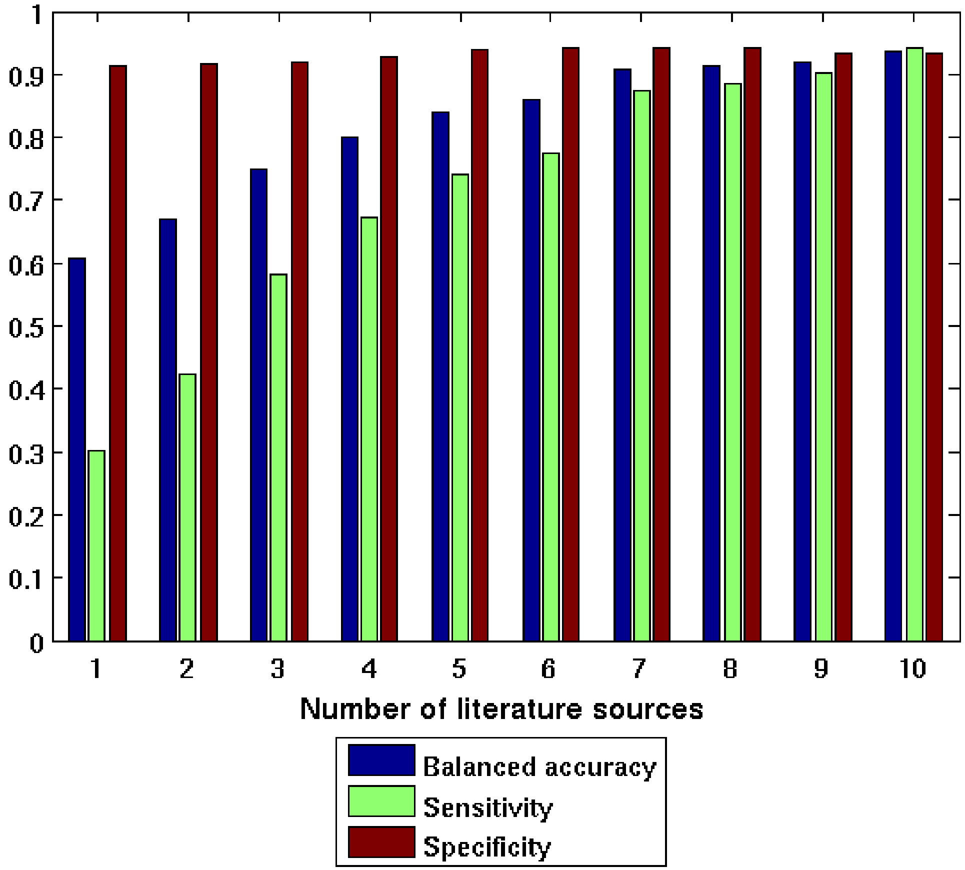
<u>Key point</u>: As greater consistency is required from literature sources, model performance improves

Consensus Model

Model with the best data

Model performance increases

# of chemicals to evaluate decreases



<u>Total Database</u>

Binders: 3961

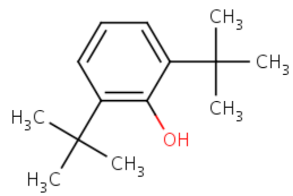
Agonists: 2494

Antagonists: 2793

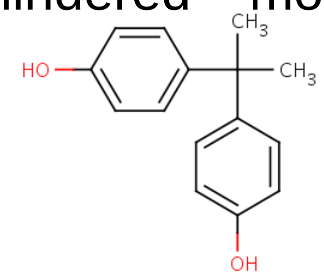
Mansouri, et al. EHP 2016

Issue with global models: Phenols are mostly predicted positive

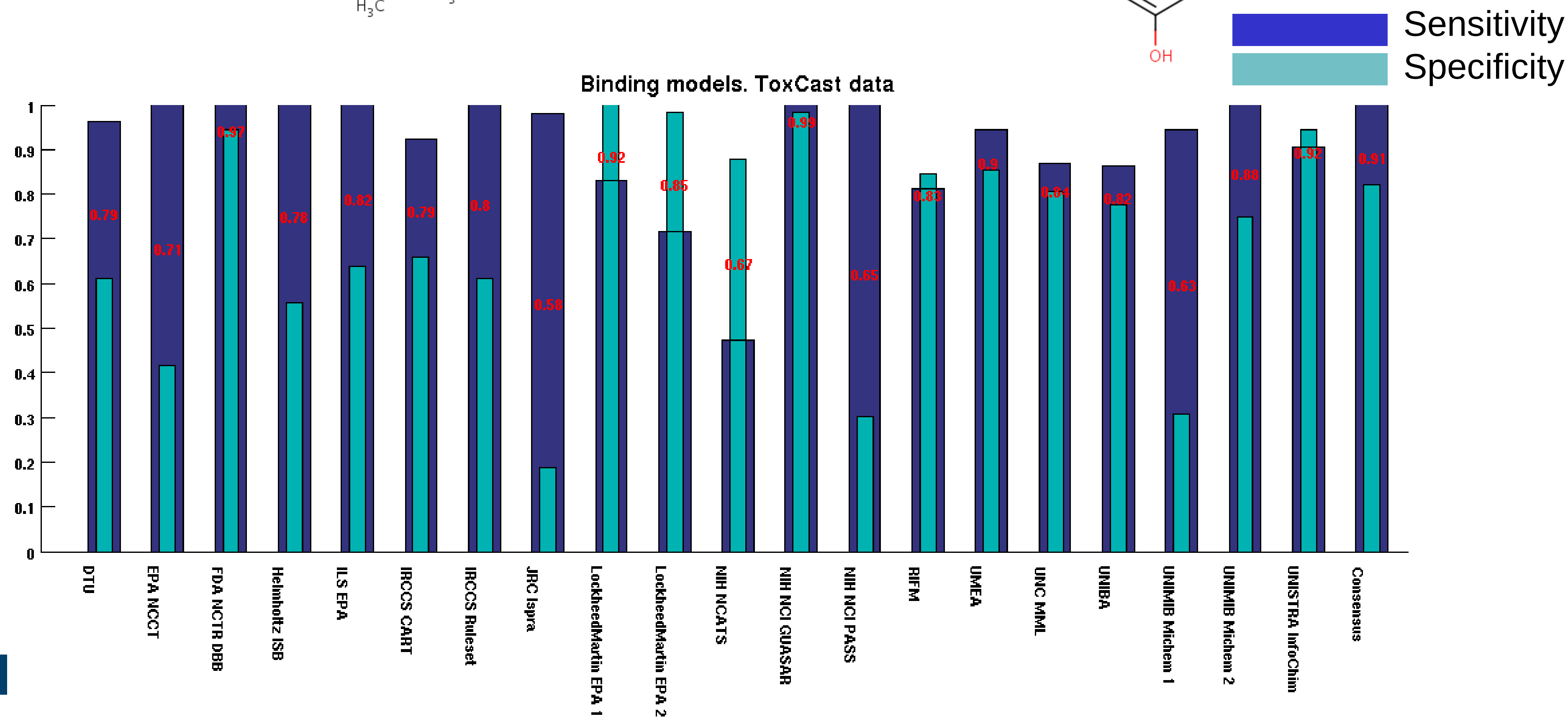
Hindered – mostly inactive



Non-hindered – mostly active



Build Local Models



Mansouri, et al. (in preparation)

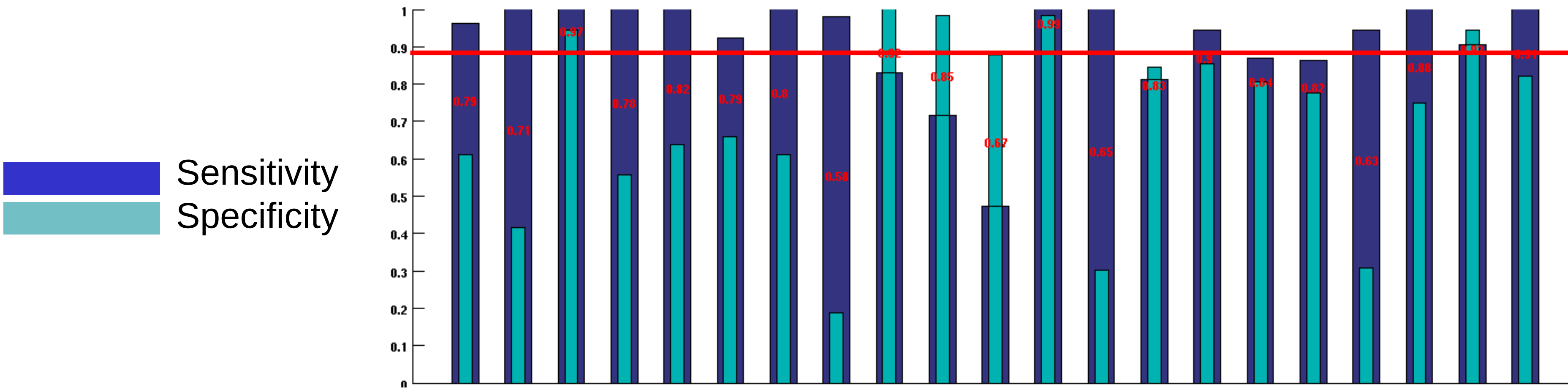
By building a local QSAR model, we can improve local accuracy

PLSDA model: 30 Descriptors, 3 Latent variables

	Training set (483)		Test set (120)
	Calibration	5-Fold CV	validation
SN	0.89	0.88	0.91
SP	0.86	0.85	0.88
BA	0.88	0.87	0.89

Build Local Models

Local model has better balanced accuracy than 17/21 global models and about same as global consensus



Mansouri, et al. (in preparation)

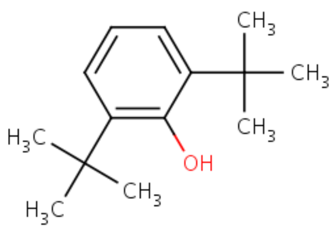
I'm finally getting to read-across!

- Need to focus locally in chemistry and bioactivity
  - Phenols / estrogen
- Need to be conscious of messiness of training and test data
  - All assays are noisy
  - And there is real biological variations between cell types, etc.
- Need to have a goal
  - Can read-across beat a "thoughtless" QSAR model?

Build Local Models

HINDERED PHENOL CASE STUDY - Health Canada and US EPA

Hindered phenols are phenols with one or more bulky functional groups ortho to the hydroxyl group.



Build Local Models

Goal: Risk assessment and categorization of 21 Hindered Phenols (HP) under the Chemicals Management Plan. One of the issues is to investigate whether particular HPs have the potential to be estrogenic or not, and if so, their relative potency using read-across and/or (Q)SAR methods.

Pradeep et al. (in prep)

EPA United States Environmental Protection Agency

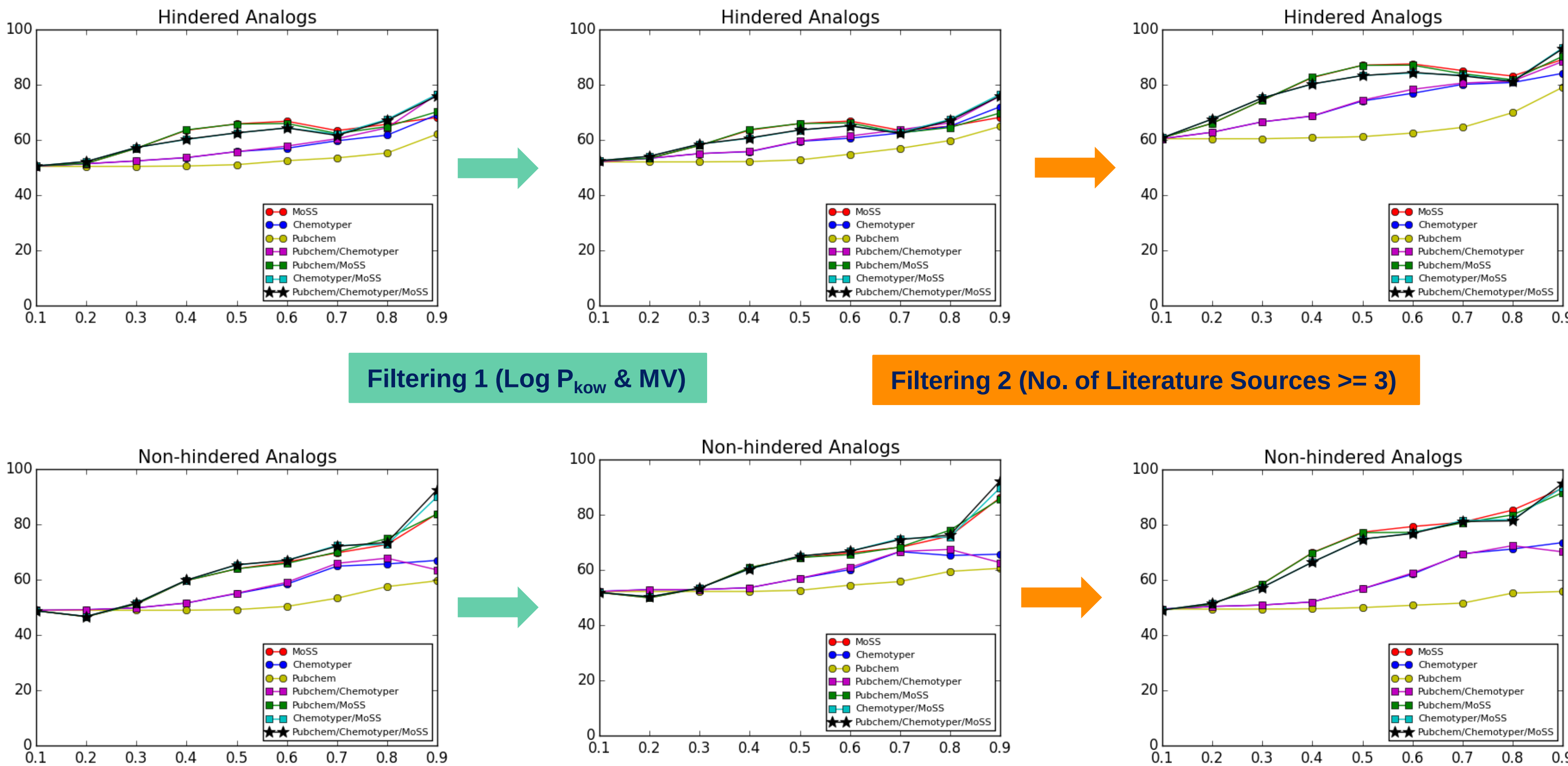
READ-ACROSS PREDICTIONS

Build Local Models

Model with the best data

Accuracy increases as

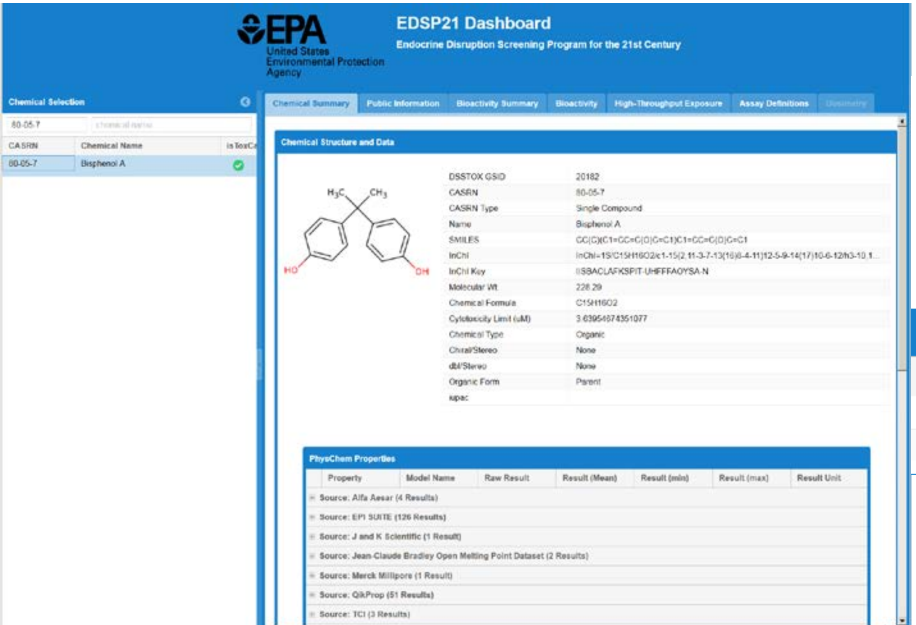
1. Better data is used in the evaluation
2. Neighbors are closer (structure and physchem)



Pradeep et al. (in prep)

Data Transparency: EDSP21 Dashboard

- Goal: To make ER and AR data easily available to all stakeholders
  - Assay-by-assays concentration-response plots
  - Model scores – AUC agonist and antagonist
  - ER QSAR calls
  - Other relevant data
- http://actor.epa.gov/edsp21



ToxCast Model Predictions

Model	Agonist AUC	Antagonist AUC
ER	0.45	0
AR	0	0.136

Consensus CERAPP QSAR ER Model Predictions

Class	Agonist (Potency Level)	Antagonist (Potency Level)	Binding (Potency Level)
from Literature	Active (Weak)	-	Active (Weak)
QSAR Consensus	Active (Weak)	Active (Strong)	Active (Weak)

Summary

- Goal is to build predictive models in the presence of noisy data

- Recognize and quantify uncertainty
- Build models on the best (most reproducible) data
- Combine multiple imperfect models together (consensus)
- Build local models where possible

Acknowledgements

Kamel Mansouri
Nicole Kleinstreuer
Eric Watt
Prachi Pradeep
Patience Browne
Grace Patlewicz
Imran Shah

NCCT Staff Scientists
Rusty Thomas
Kevin Crofton
Keith Houck
Ann Richard
Richard Judson
Tom Knudsen
Matt Martin
Grace Patlewicz
Woody Setzer
John Wambaugh
Tony Williams
Steve Simmons
Chris Grulke
Jim Rabinowitz

NCCT
Nancy Baker
Jeff Edwards
Dayne Filer
Parth Kothiya
Doris Smith
Jamey Vail
Sean Watford
Indira Thillainadarajah

NCCT Postdocs
Todor Antonijevic
Audrey Bone
Kristin Connors
Danica DeGroot
Jeremy Fitzpatrick
Jason Harris
Dustin Kapraun
Agnes Karmaus
Max Leung
Kamel Mansouri
LyLy Pham
Prachi Pradeep
Caroline Ring
Eric Watt

NIH/NCATS
Menghang Xia
Ruili Huang
Anton Simeonov

NTP
Warren Casey
Nicole Kleinstreuer
Mike Devito
Dan Zang