Consensus Models to Predict Endocrine Disruption for Human-Exposure Chemicals

Kamel Mansouri

Richard Judson

ScitoVation LLC
NCCT, U.S. EPA
RTP, NC, USA

13 June 2017

Background

- U.S. Congress mandated that the EPA screen chemicals for their potential to be endocrine disruptors
- Led to development of the Endocrine Disruptor Screening Program (EDSP)
- Initial focus was on environmental estrogens, but program expanded to include androgens and thyroid pathway disruptors

Endocrine Disruptor Screening Program

- Concern over environmental chemical disruption of endocrine hormone signaling
- Congressionally mandated, multiple EDSP testing tiers (11 tests in Tier 1)
- EDSP Tier 1 Testing: for the purposes of prioritization and screening, identify chemicals with the potential to disrupt estrogen, androgen, or thyroid hormone receptor signaling.
- There is a mismatch between resources needed for EDSP Tier 1 testing and the number of chemicals to be tested
- **New Approach**: **EDSP + Tox21 = EDSP21**
  - Pathway-based models
  - Multiple high-throughput in vitro assays
  - Validate to replace selected Tier 1 screening assays

EDSP Chemicals

- EDSP Legislation contained in:
  - FIFRA: Federal Insecticide, Fungicide, Rodenticide Act
  - SDWA: Safe Drinking Water Act
- Chemicals:
  - All pesticide ingredients (actives and inerts)
  - Chemicals likely to be found in drinking water to which a significant population can be exposed
- Total EDSP Chemical universe is ~10,000
- Subsequent filters brings this to about 5,000 to be tested

Problem statement

- EDSP Consists of Tier 1 and Tier 2 tests
- Tier 1 is a battery of 11 in vitro and in vivo assays
- Cost ~$1,000,000 per chemical
- Throughput is ~50 chemicals / year
- Total cost of Tier 1 is billions of dollars and will take 100 years at the current rate
- Need pre-tier 1 filter
- Use combination of structure modeling tools and high-throughput screening “EDSP21”

Tox21/ToxCast

- Tox21: Federal consortium including EPA, FDA,, NCGC,NCATS, NTP, NIEHS
  - ~10k chemicals x 60 assays
- ToxCast: EPA's Toxicity Forecaster
  - ~2k chemicals x 800 assays
- High-throughput assays for these targets or pathways
- Develop predictive systems models
- Use predictive models (qualitative):
  - Prioritize chemicals for targeted testing
  - Suggest / distinguish possible AOPs
- Use predictive models (quantitative):
  - Screen chemicals for hazard
  - Green chemistry design

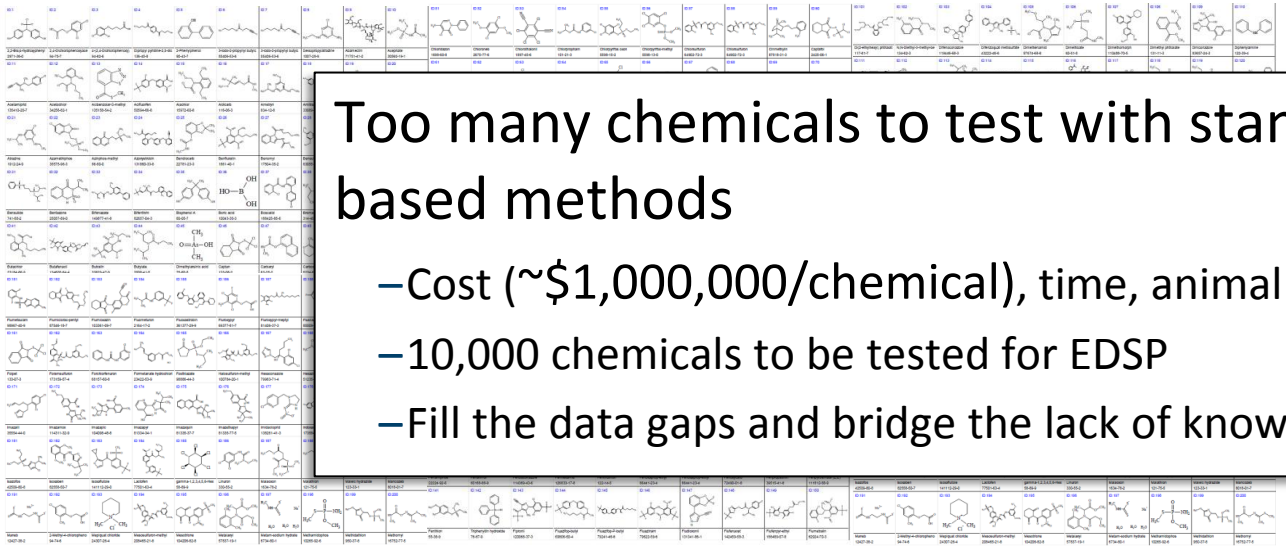
General goals

- Use structure-based models to predict ER + AR activity for all of EDSP Universe and aid in prioritization for EDSP Tier 1

- Because models are relatively easy to run on large numbers of chemicals, extend to all chemicals with likely human exposure

- Chemicals with significant evidence of ER + AR activity can be queued further testing

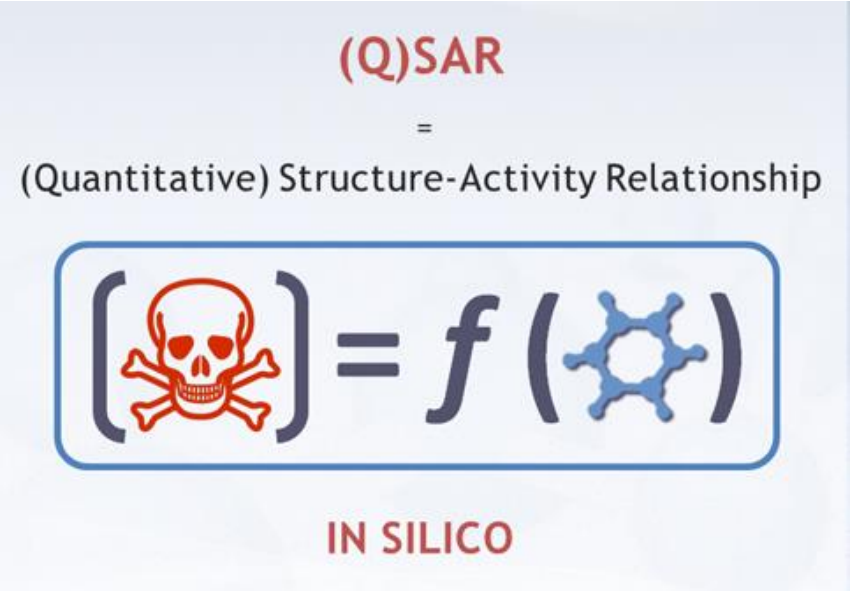
Computational Toxicology



Too many chemicals to test with standard animal-based methods

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

Alternative



Quantitative Structure Activity/Property Relationships (QSAR/QSPR)

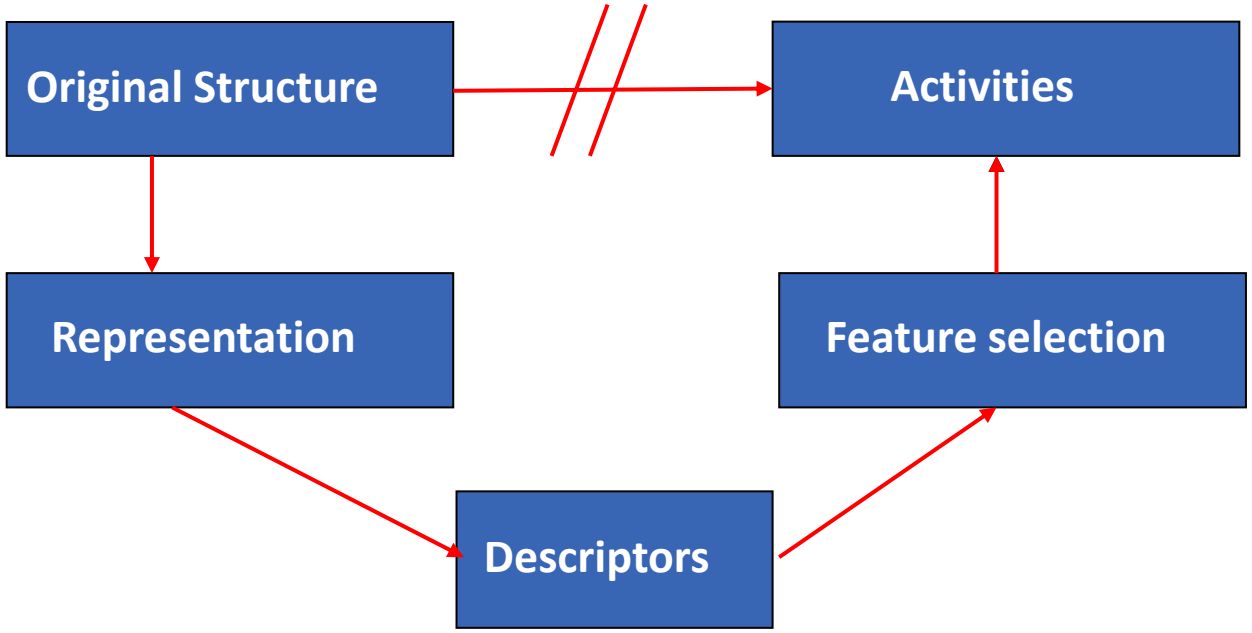
Congenericity principle: QSARs correlate, within congeneric series of compounds, their chemical or biological activities, either with certain structural features or with atomic, group or molecular descriptors.

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

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

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

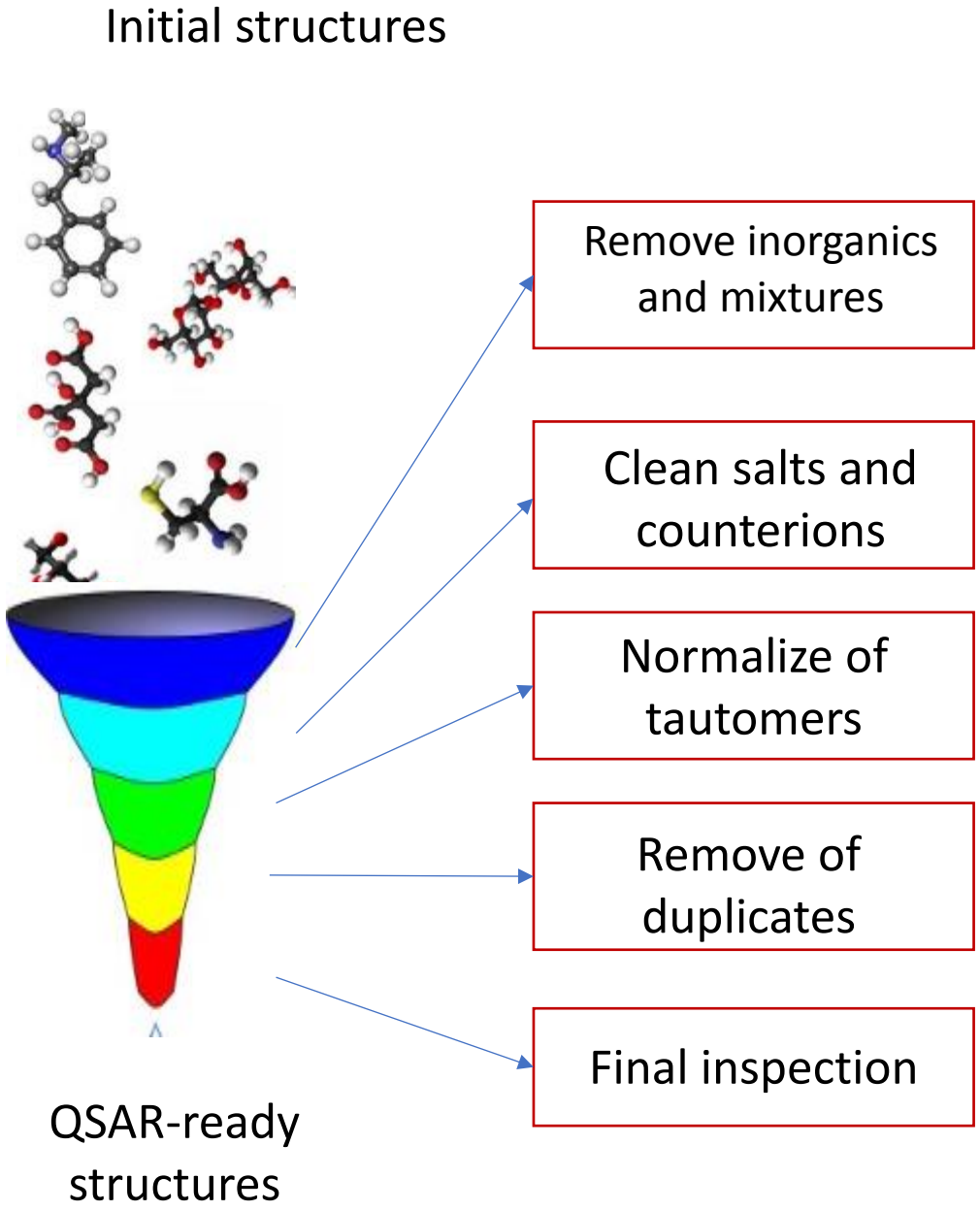
b_i - ***fitted parameters***



Development of a QSAR model

- Curation of experimental data (Data may be noisy and limits prediction accuracy)
- Preparation of training and test sets
- Calculation of an initial set of descriptors
- Selection of a mathematical method
- Variable selection technique
- Validation of the model's predictive ability
- Define the Applicability Domain

Structure standardization



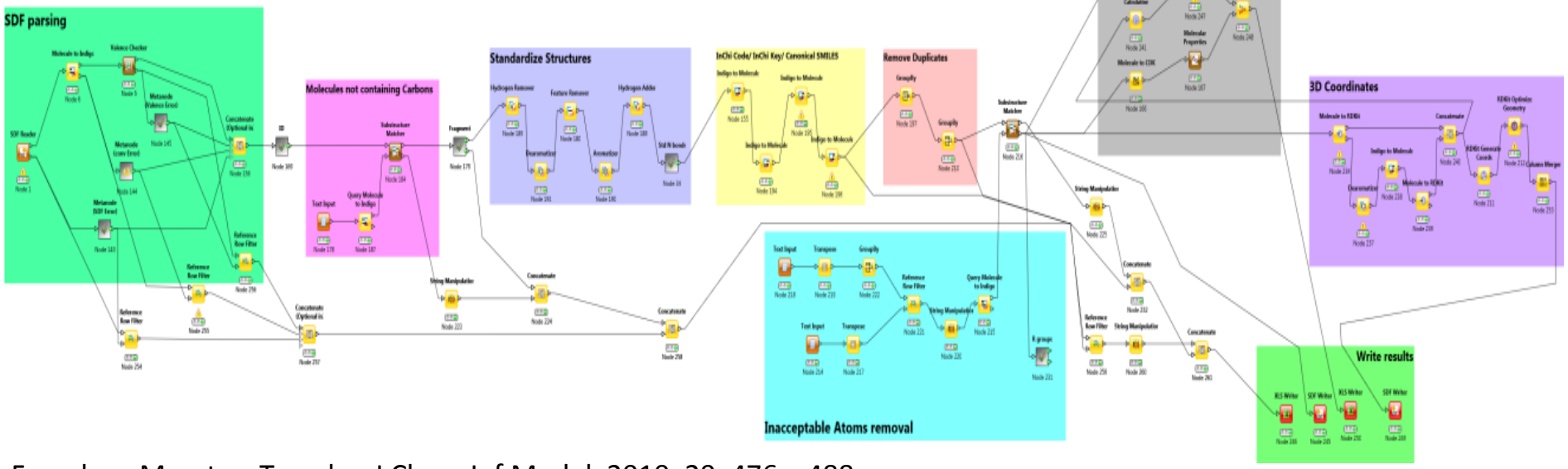
KNIME workflow

Aim of the workflow:

- Combine (<u>not reproduce</u>) different procedures and ideas
- Minimize the differences between the structures used for prediction by different groups
- Produce a flexible free and open source workflow to be shared

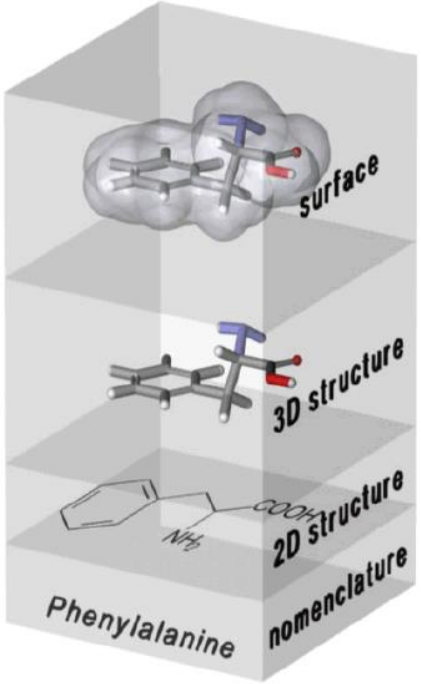
Indigo

RDKit
Open-Source Cheminformatics and Machine Learning



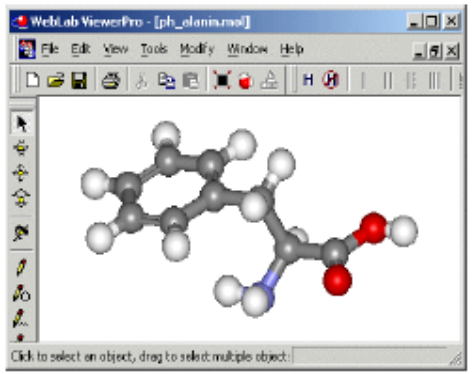
Fourches, Muratov, Tropsha. J Chem Inf Model, 2010, 29, 476 – 488
Wedebye, Niemelä, Nikolov, Dybdahl, Danish EPA Environmental Project No. 1503, 2013

Molecular structures in the computer

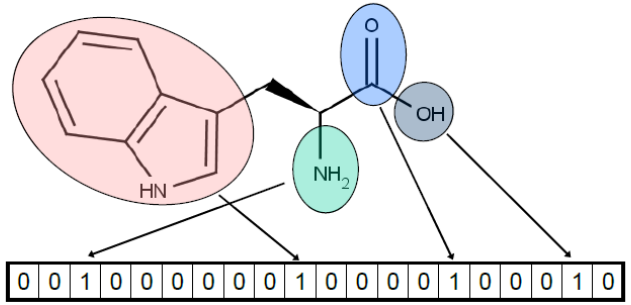



```
C9H11NO2
DAtclserve10160209553D 0   0.000

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



Bitstrings in databases

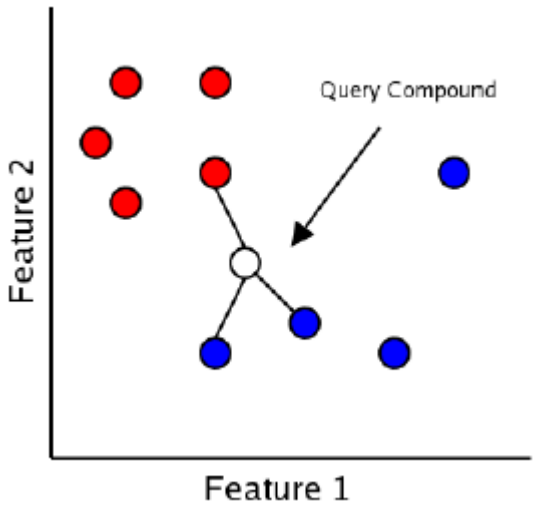


<u>Fragmental keys & fingerprints</u>

- substructural search
- read-across
- similarity search

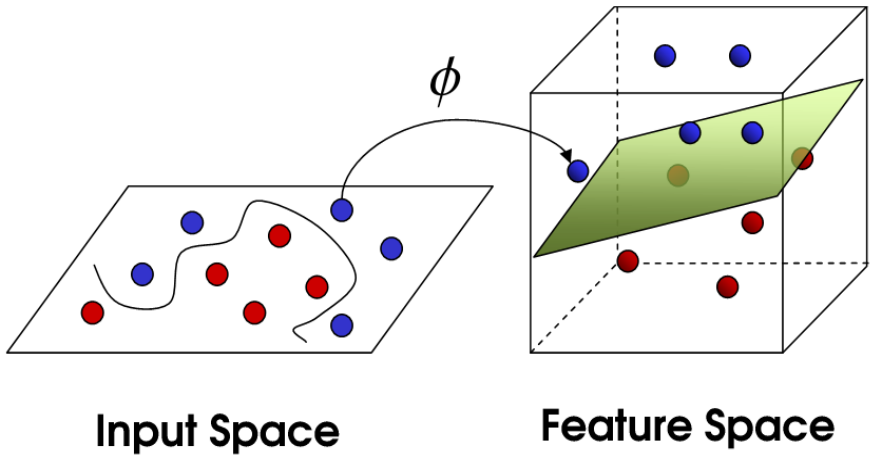
Classification methods

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



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

- **SVM: Support Vector Machines**



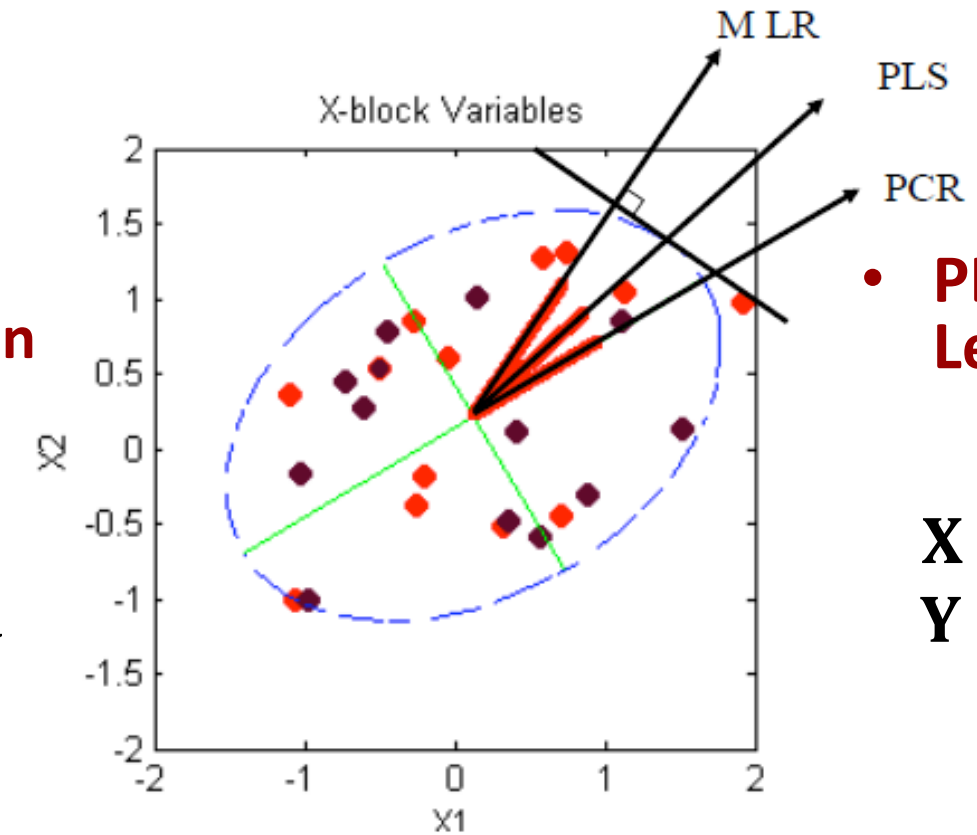
Kernel function maximizing the margin between the classes

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

Regression methods

- **MLR: Multiple Linear Regression**

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



- **PLS: Partial Least Squares**

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

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

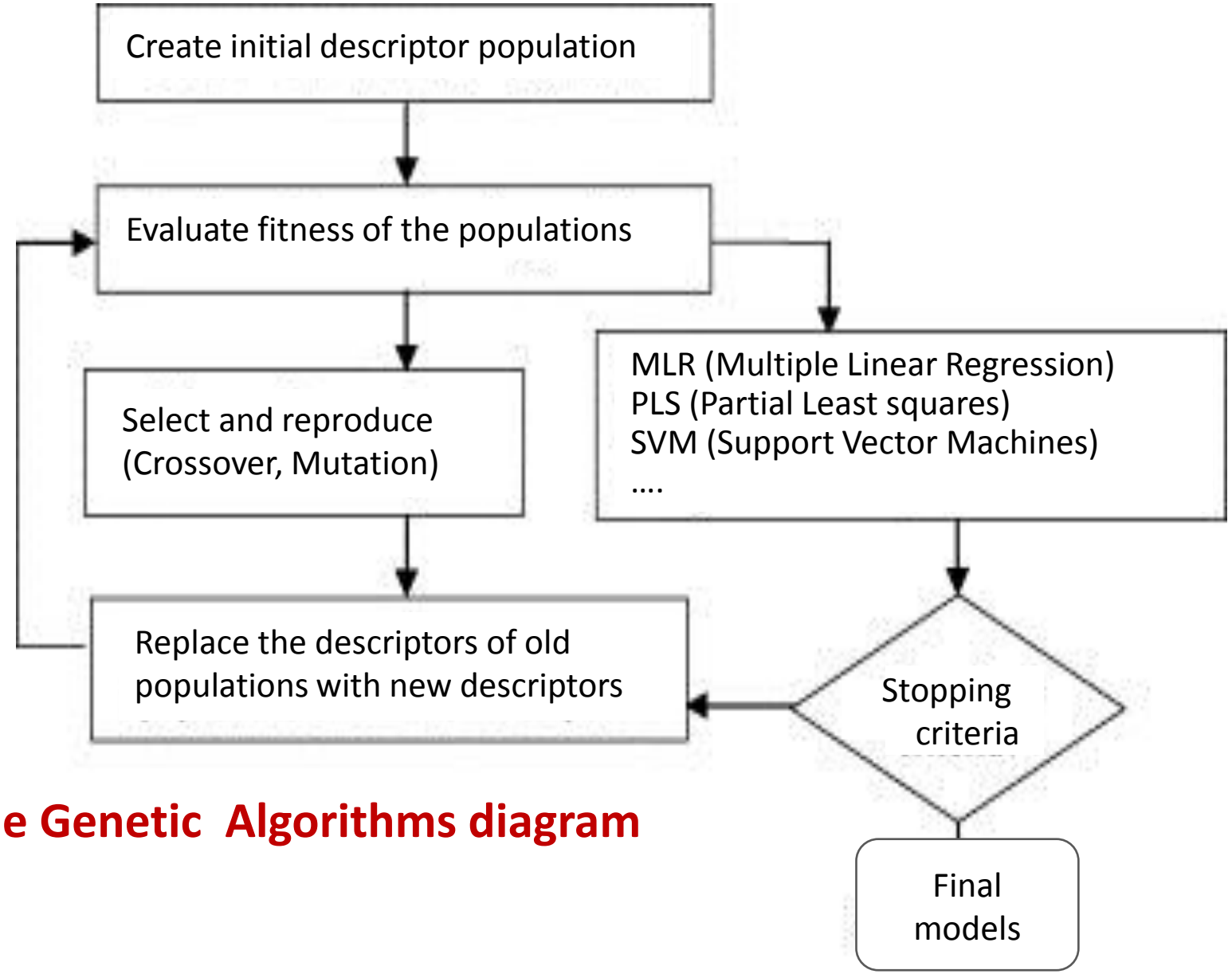
Other methods: Artificial Neural Networks (ANN), Random Forest, LASSO, PCR…

Variable selection procedure

- **Many more descriptors than chemicals**

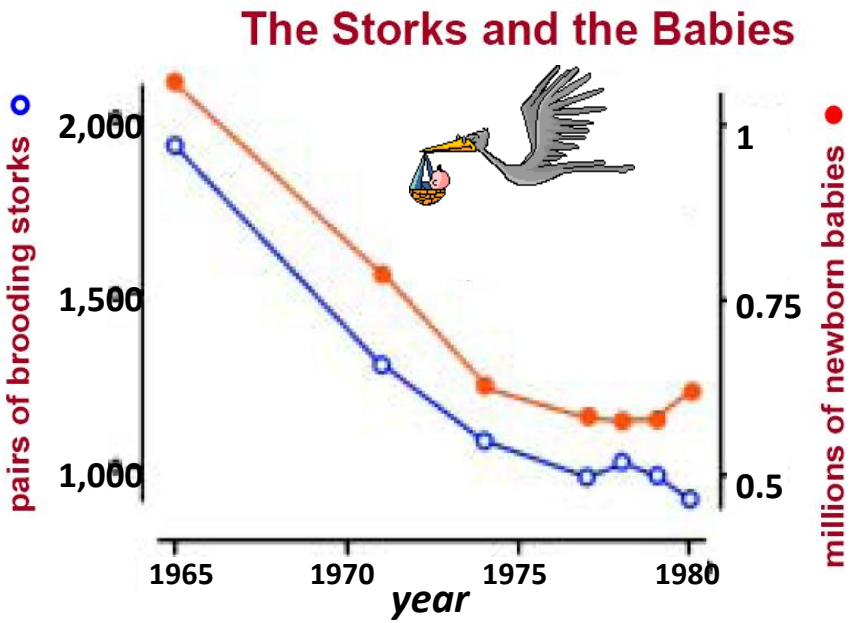
- **Many irrelevant descriptors**

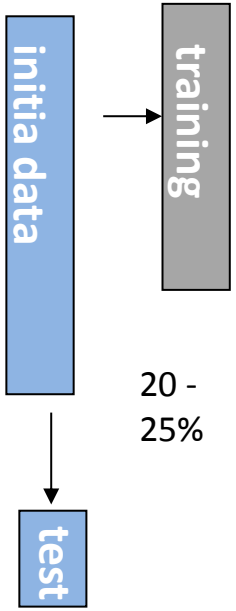
Only the most important descriptors are selected

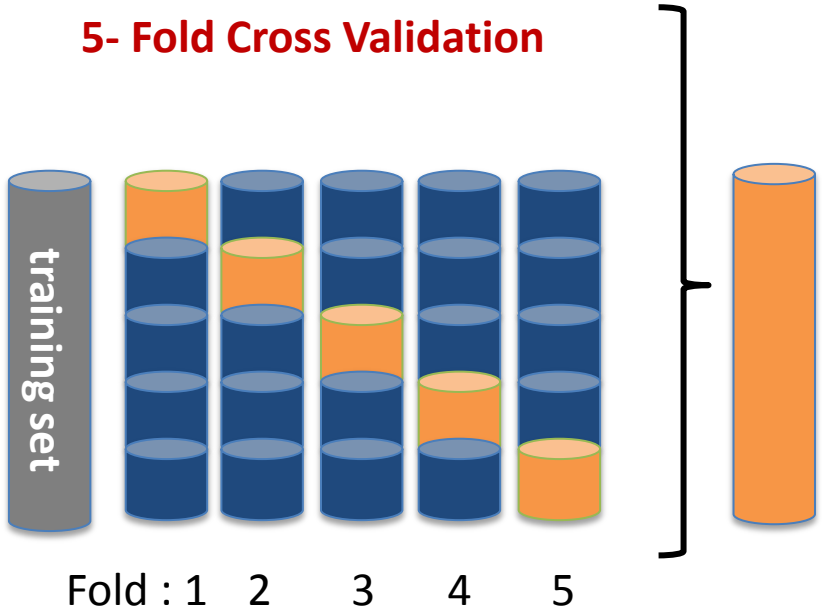


The Genetic Algorithms diagram

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







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

H. Sies, Nature 332, 495 (1988)

CERRAP : Collaborative Estrogen Receptor Activity Prediction Project

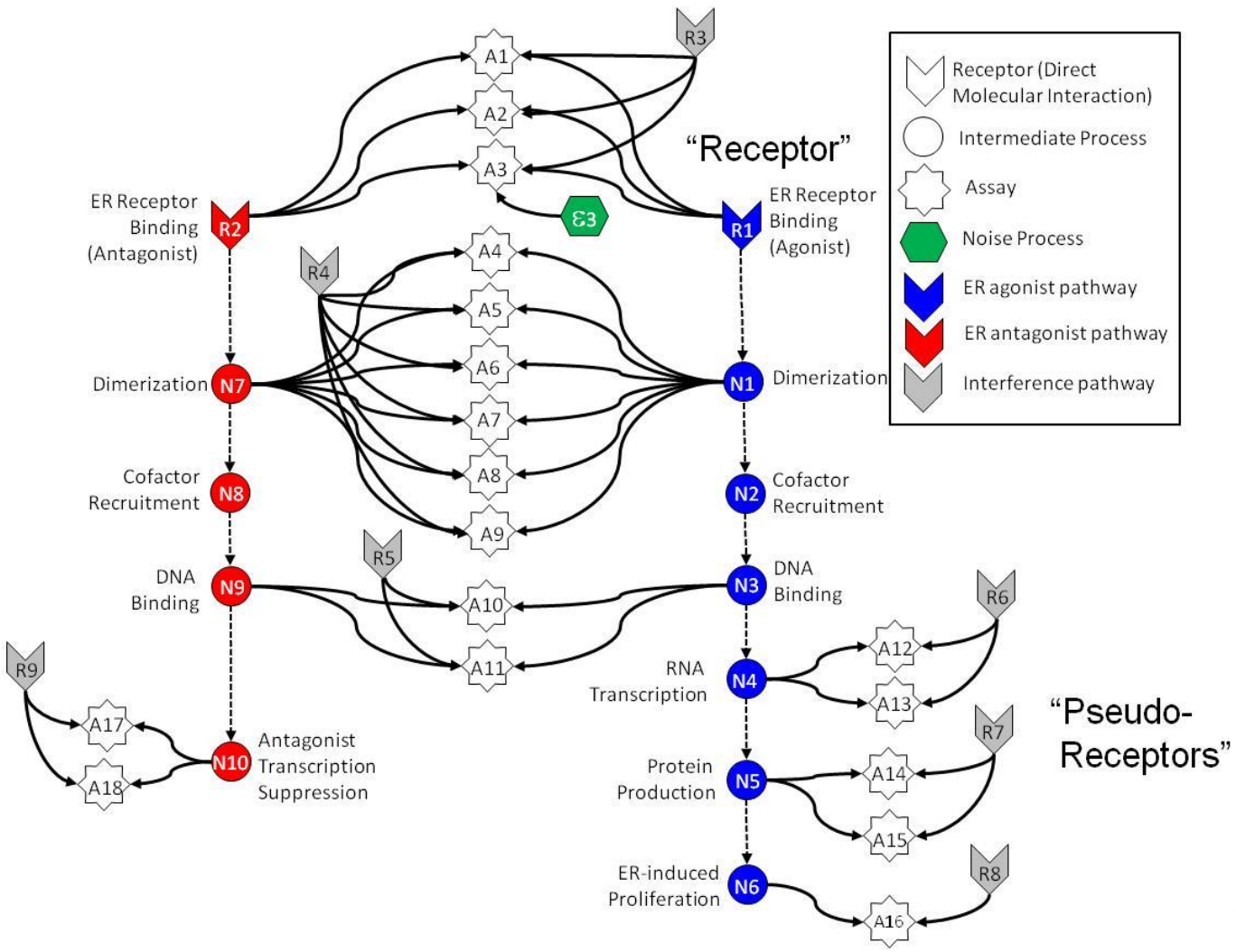
40 scientists, 17 groups

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

Plan of the project

1: Structures curation	- Collect chemical structures from different sources - Design and document a workflow for structure cleaning - Deliver the QSAR-ready training set and prediction set
2: Experimental data preparation	- Collect and clean experimental data for the evaluation set - Define a strategy to evaluate the models separately
3: Modeling & predictions	- Train/refine the models based on the training set - Deliver predictions and applicability domains for evaluation
4: Model evaluation	- Analyze the training and evaluation datasets - Evaluate the predictions of each model separately
5: Consensus strategy	- Define a score for each model based on the evaluation step - Define a weighting scheme from the scores
6: Consensus modeling & validation	- Combine the predictions based on the weighting scheme - Validate the consensus model using an external dataset.

Tox21/ToxCast ER Pathway Model



ToxCast High Throughput Screening ER assays		
Assay Name	Biological Process	Assay #
NVS_NR_bER	receptor binding	1
NVS_NR_hER	receptor binding	2
NVS_NR_mERa	receptor binding	3
OT_ER_ERaERa_0480	protein complementation	4
OT_ER_ERaERa_1440	protein complementation	5
OT_ER_ERaERb_0480	protein complementation	6
OT_ER_ERaERb_1440	protein complementation	7
OT_ER_ERbERb_0480	protein complementation	8
OT_ER_ERbERb_1440	protein complementation	9
OT_ERa_EREGFP_0120	gene expression	10
OT_ERa_EREGFP_0480	gene expression	11
ATG_ERa_TRANS_up	mRNA induction	12
ATG_ERE_CIS_up	mRNA induction	13
Tox21_ERa_BLA_Agonist_ratio	gene expression	14
Tox21_ERa_LUC_BG1_Agonist	gene expression	15
ACEA_T47D_80hr_Positive	cell proliferation	16
Tox21_ERa_BLA_Antag_ratio	gene expression	17
Tox21_ERa_LUC_BG1_Antag	gene expression	18

Judson et al Toxicol. Sci. (2015) 148 (1): 137-154. doi: 10.1093/toxsci/kfv168

Computational Model

$$A_i = \sum_j F_{ij} R_j$$

$\boldsymbol{A_i}$ is the efficacy of the assay at a given concentration
$\boldsymbol{R_j}$ is the "true" efficacy which is unobservable
$\boldsymbol{F}$ links receptors to assays

$$\varepsilon^2 = \sum_i (A_i^{pred} - A_i^{meas})^2 + penalty(\vec{R})$$

Solve a constrained least-squares problem to minimize difference between the measured and predicted assay values

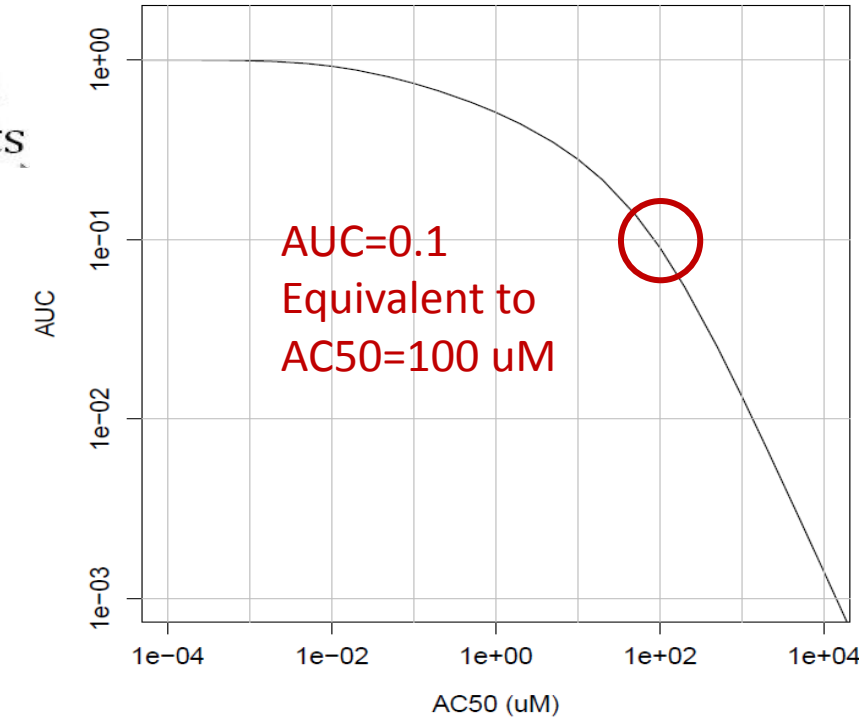
$$A_i^{pred} \in [1,0]$$

$$penalty(\vec{R}) = \alpha \frac{SR^2}{SR^2 + SR_0{}^2}$$

Penalty enforces physical assumption that chemical will not hit many targets simultaneously

$$AUC_j = \frac{1}{N_{conc}} \sum_{i=1}^{N_{conc}} sign(slope) \times R_j(conc_i)$$

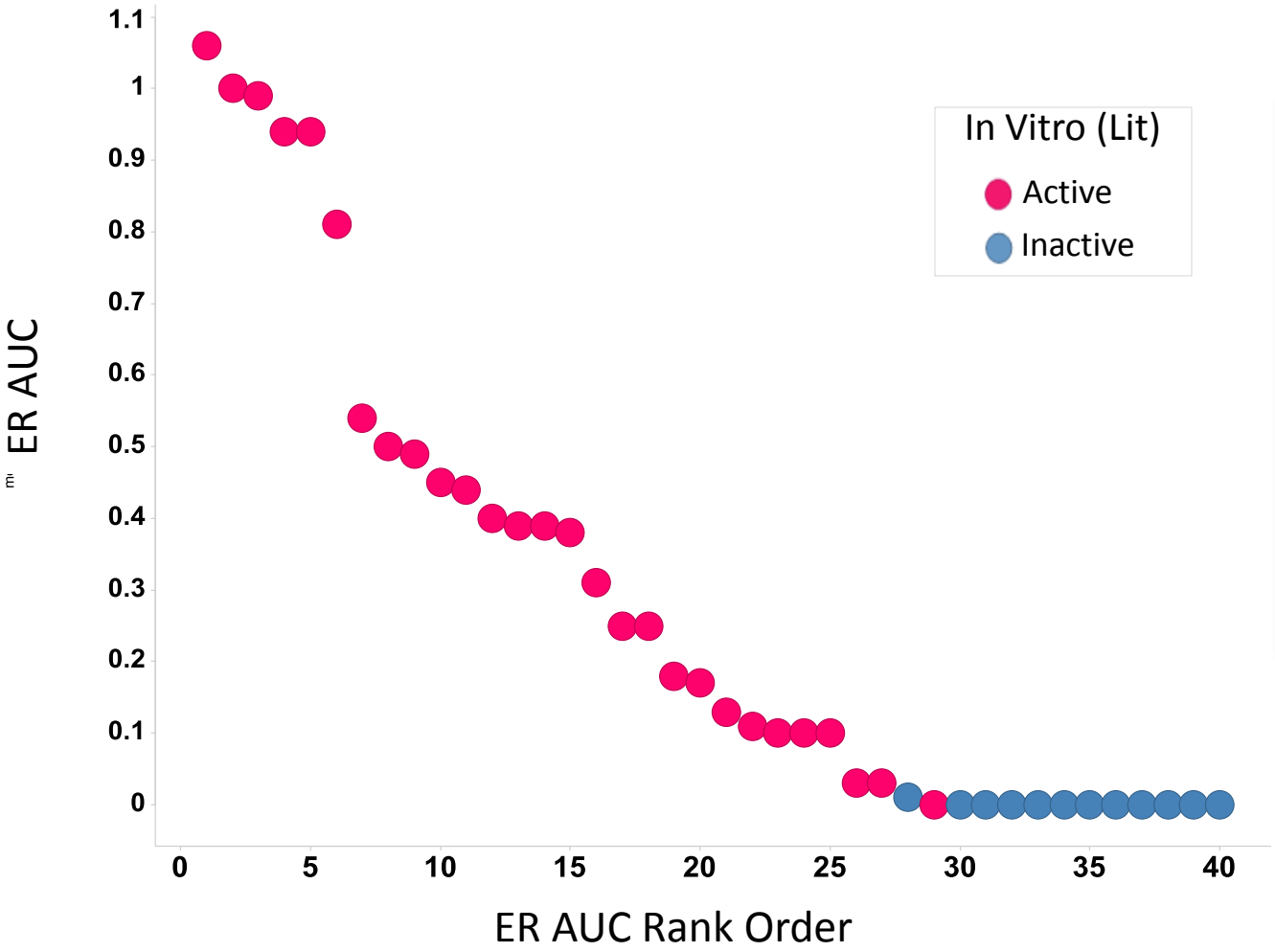
AUC Summarizes results



Judson et al Toxicol. Sci. (2015) 148 (1): 137-154. doi: 10.1093/toxsci/kfv168

ER Model Performance

In Vitro Reference Chemicals

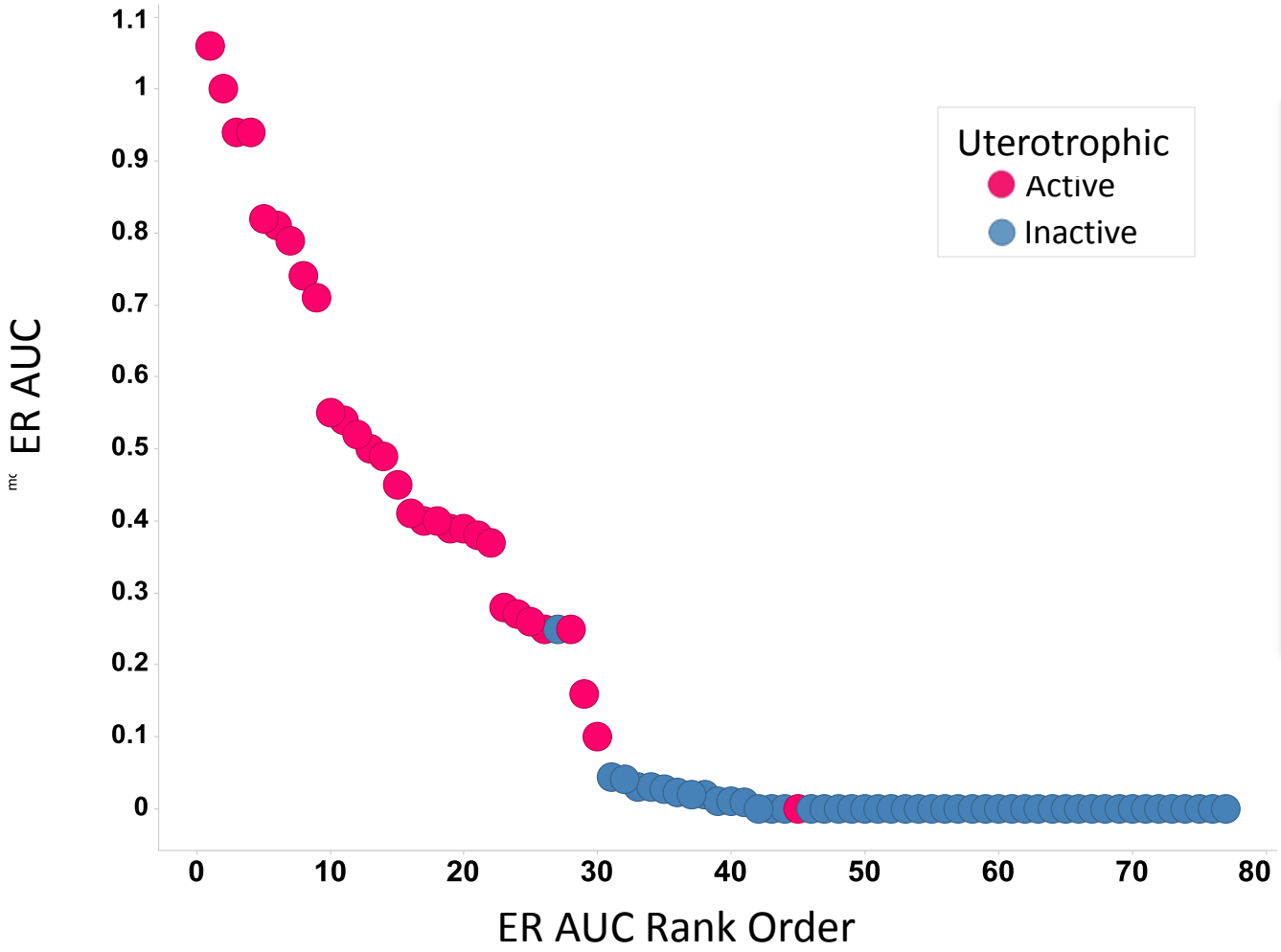


True Positive	25
True Negative	12
False Positive	0
False Negative	3
Accuracy	**0.95**
Sensitivity	**0.89**
Specificity	**1.00**

Judson et al. 2015 Tox Sci

ER Model Performance

In Vivo Reference Chemicals



True Positive	29
True Negative	46
False Positive	1
False Negative	1
Accuracy	**0.97**
Sensitivity	**0.97**
Specificity	**0.97**

Browne et al. 2015 ES&T

Chemicals for Prediction:
The Human Exposure Universe

- EDSP Universe (10K)
- Chemicals with known use (40K) (CPCat & ACToR)
- Canadian Domestic Substances List (DSL) (23K)
- EPA DSSTox – structures of EPA/FDA interest (15K)
- ToxCast and Tox21 (In vitro ER data) (8K)

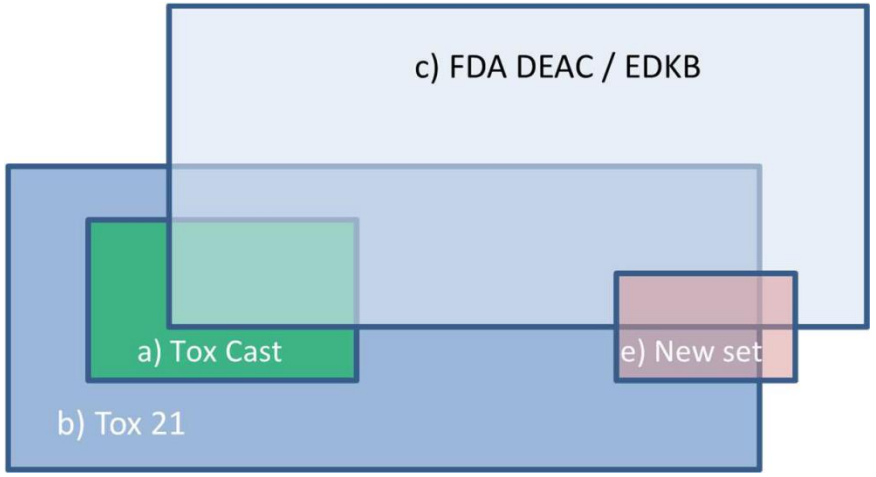
→ **~55k to ~32K unique set of structures**

- Training set (ToxCast): 1677 Chemicals
- Prediction Set: 32464 Chemicals

Experimental data for evaluation set

a) Tox21, ~8000 chemicals in 4 assays;
b) FDA EDKB database of ~8000 chemicals from the literature;
c) METI database, ~2000 chemicals;
d) ChEMBL database, ~2000 chemicals.

60,000 entries for ~15,000 chemicals



CERAPP models

- Training set (ToxCast): 1677 Chemicals
- Prediction Set: 32464 Chemicals

Models received:

- Classification / Qualitative:
  - Binding: **22 models**
  - Agonists: **11 models**
  - Antagonists: **9 models**
- Regression / Quantitative:
  - Binding: **3 models**
  - Agonists: **3 models**
  - Antagonists: **2 models**

Evaluation procedure:

- On the EPA training set (1677)
- On the full evaluation set (~7k)
- Evaluation set with multi-sources
- Remove “VeryWeak” & ambiguous
- Remove chemicals outside the AD

→ **Score functions & weights for consensus predictions**

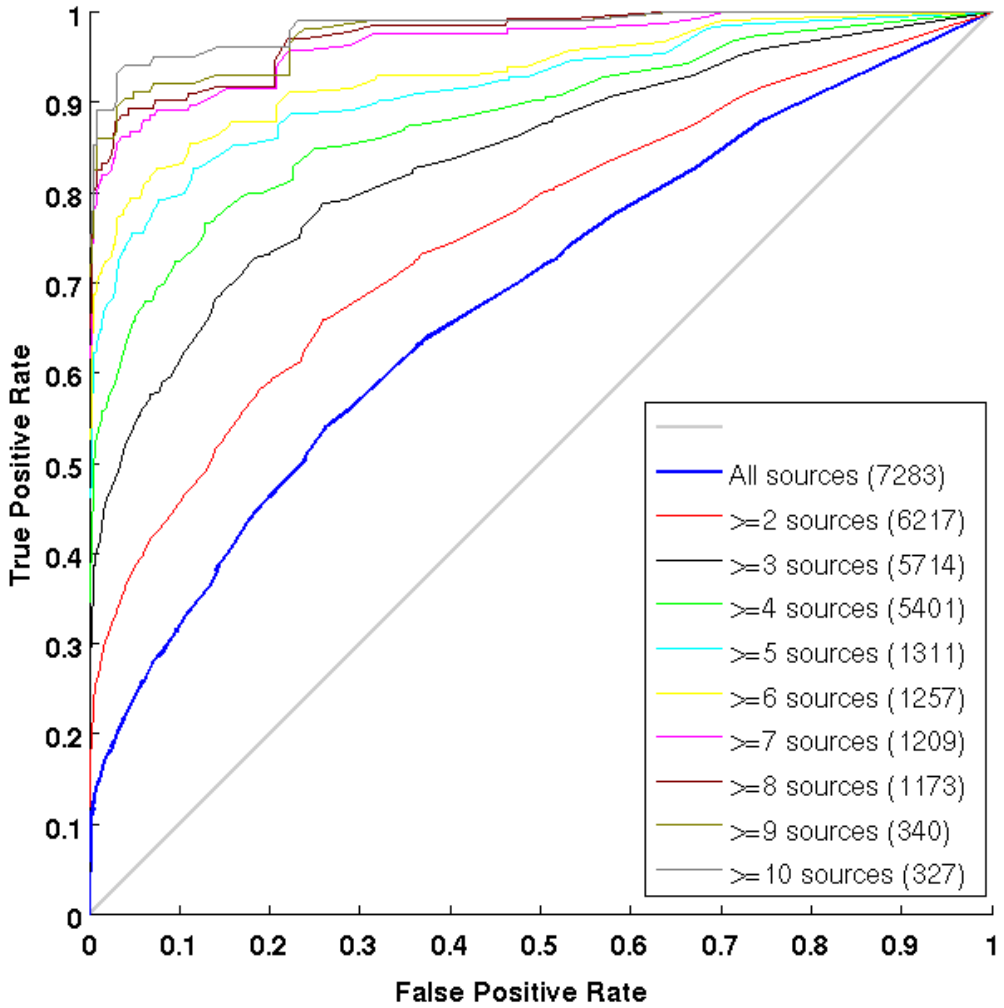
Consensus Qualitative Accuracy

Prediction Accuracy Strongly Depends on Data Quality

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

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

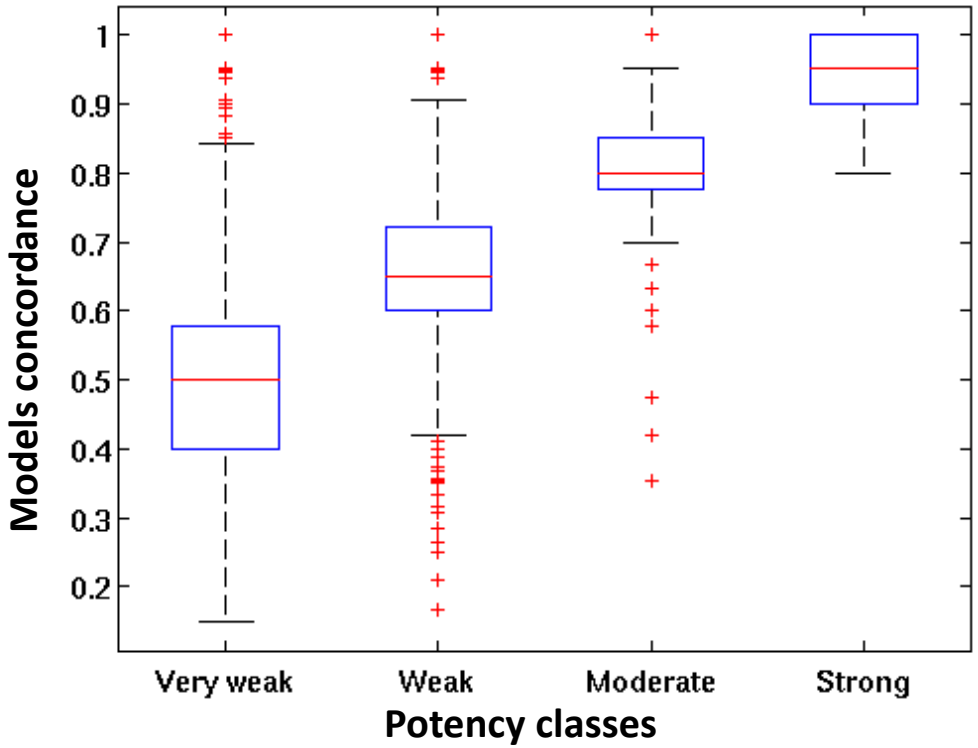
	ToxCast data	Literature data (All: 7283)	Literature data (>6 sources: 1209)
Sensitivity	0.93	0.30	0.87
Specificity	0.97	0.91	0.94
Balanced accuracy	0.95	0.61	0.91



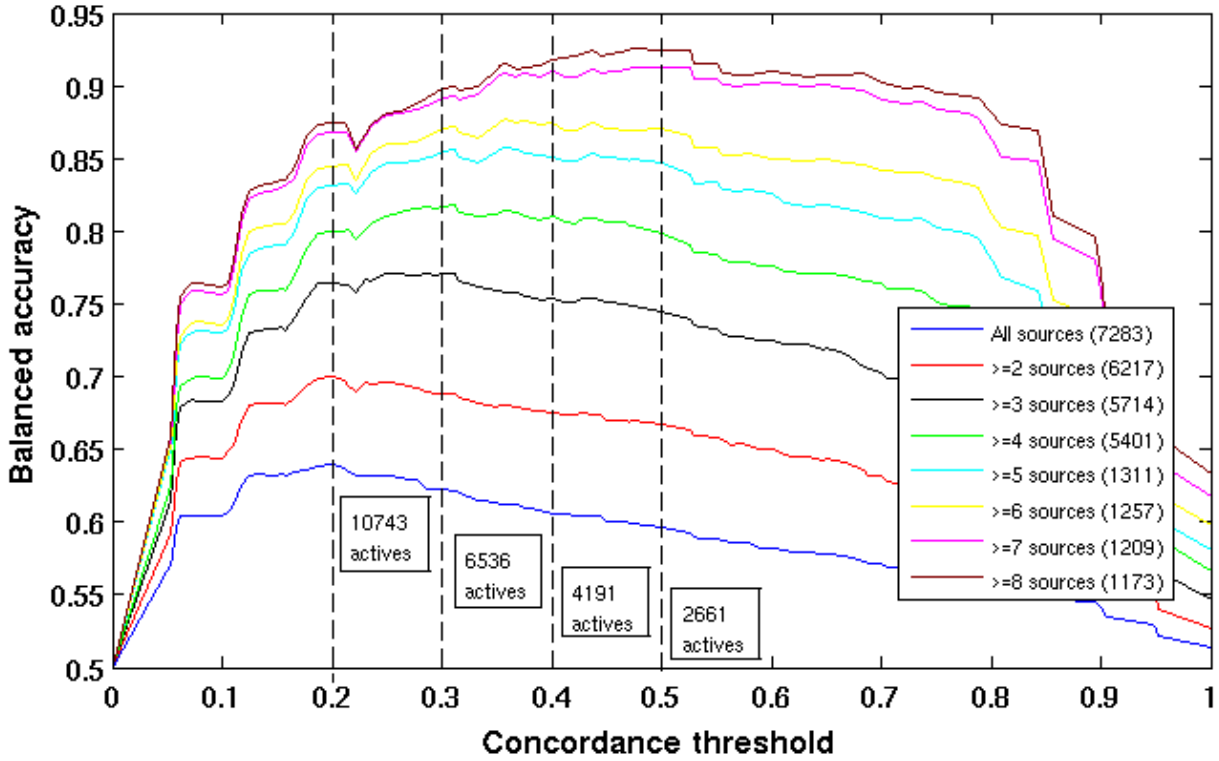
ROC curve of the external validation set (literature)

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

Consensus Quantitative Accuracy



Box plot of the active classes of the consensus model.



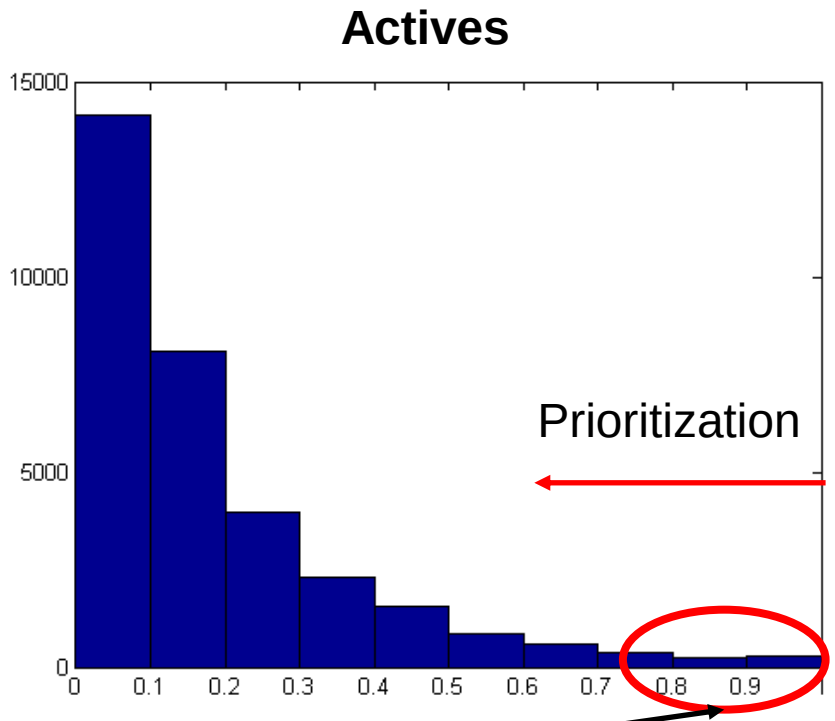
Variation of the balanced accuracy with positive concordance thresholds

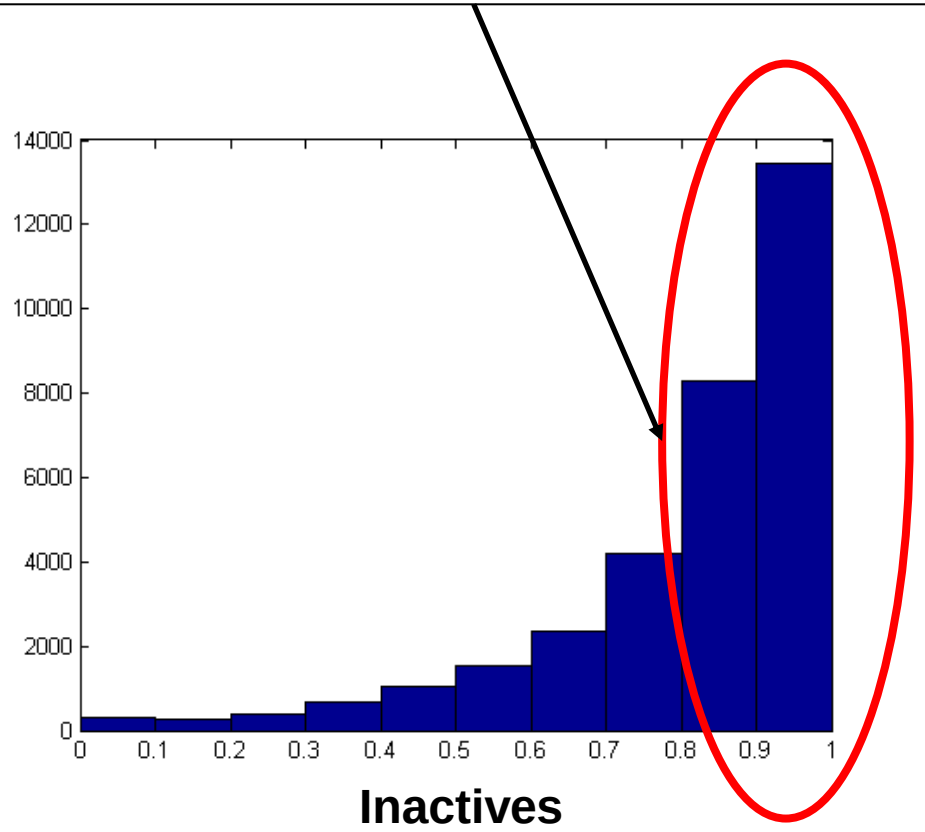
- positive concordance < 0.6 => Potency class= **Very weak**
- 0.6=<positive concordance<0.75 => Potency class= **Weak**
- 0.75=<positive concordance<0.9 => Potency class= **Moderate**
- positive concordance>=0.9 => Potency class= **Strong**

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

Concordance of the qualitative models

Most models predict most chemicals as inactive





Only 757 chemicals have >75% positive concordance

⇨ Only a small fraction of chemicals require further testing!

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

ehp ENVIRONMENTAL HEALTH PERSPECTIVES

Mansouri et al. (2016)
DOI:10.1289/ehp.1510267

Articles | Collections | Authors | *EHP* 中文版 | Careers and Funding | E-Mail Alerts | About *EHP*

RESEARCH ARTICLE VOLUME 124 | ISSUE 7 | JULY 2016

Environ Health Perspect; DOI:10.1289/ehp.1510267

CERAPP: Collaborative Estrogen Receptor Activity Prediction Project

Kamel Mansouri,[1,2] Ahmed Abdelaziz,[3] Aleksandra Rybacka,[4] Alessandra Roncaglioni,[5] Alexander Tropsha,[6] Alexandre Varnek,[7] Alexey Zakharov,[8] Andrew Worth,[9] Ann M. Richard,[1] Christopher M. Grulke,[1] Daniela Trisciuzzi,[10] Denis Fourches,[6] Dragos Horvath,[7] Emilio Benfenati,[5] Eugene Muratov,[6] Eva Bay Wedebye,[11] Francesca Grisoni,[12] Giuseppe F. Mangiatordi,[10] Giuseppina M. Incisivo,[5] Huixiao Hong,[13] Hui W. Ng,[13] Igor V. Tetko,[3,14] Ilya Balabin,[15] Jayaram Kancherla,[1] Jie Shen,[16] Julien Burton,[9] Marc Nicklaus,[8] Matteo Cassotti,[12] Nikolai G. Nikolov,[11] Orazio Nicolotti,[10] Patrik L. Andersson,[4] Qingda Zang,[17] Regina Politi,[6] Richard D. Beger,[18] Roberto Todeschini,[12] Ruili Huang,[19] Sherif Farag,[6] Sine A. Rosenberg,[11] Svetoslav Slavov,[17] Xin Hu,[19] and Richard S. Judson[1]

Author Affiliations open

PDF Version (686 KB)

Abstract | About This Article | Supplemental Material

Background: Humans are exposed to thousands of man-made chemicals in the environment. Some chemicals mimic natural endocrine hormones and, thus, have the potential to be endocrine disruptors. Most of these chemicals have never been tested for their ability to interact with the estrogen receptor (ER). Risk assessors need tools to prioritize chemicals for

Title | Add | More | 1–20 | Cited by | Year

CERAPP: Collaborative estrogen receptor activity prediction project — 37 — 2016
K Mansouri, A Abdelaziz, A Rybacka, A Roncaglioni, A Tropsha, A Varnek, ...
Environmental health perspectives 124 (7), 1023

A renaissance of neural networks in drug discovery
II Baskin, D Winkler, IV Tetko - Expert opinion on drug discovery, 2016 - Taylor & Francis
ABSTRACT Introduction: Neural networks are becoming a very popular method for solving machine learning and artificial intelligence problems. The variety of neural network types and their application to drug discovery requires expert knowledge to choose the most
Cited by 7 Web of Science: 3 Cite Save More

ToxCast chemical landscape: Paving the road to 21st century toxicology
AM Richard, RS Judson, KA Houck... - Chemical research in ..., 2016 - ACS Publications
The US Environmental Protection Agency's (EPA) ToxCast program is testing a large library of Agency-relevant chemicals using in vitro high-throughput screening (HTS) approaches to support the development of improved toxicity prediction models. Launched in 2007, Phase I
Cited by 6 Cite Saved More

[HTML] Phytoestrogens and Mycoestrogens Induce Signature Structure Dynamics Changes on Estrogen Receptor α
X Chen, U Uzuner, M Li, W Shi, JS Yuan... - International Journal of ..., 2016 - mdpi.com
Endocrine disrupters include a broad spectrum of chemicals such as industrial chemicals, natural estrogens and androgens, synthetic estrogens and androgens. Phytoestrogens are widely present in diet and food supplements; mycoestrogens are frequently found in grains.
Cite Save More

Identifying known unknowns using the US EPA's CompTox Chemistry Dashboard
AD McEachran, JR Sobus, AJ Williams - Analytical and Bioanalytical ..., 2016 - Springer
Abstract Chemical features observed using high-resolution mass spectrometry can be tentatively identified using online chemical reference databases by searching molecular formulae and monoisotopic masses and then rank-ordering of the hits using appropriate
Cite Save More

Public (Q) SAR Services, Integrated Modeling Environments, and Model Repositories on the Web: State of the Art and Perspectives for Future Development
IV Tetko, U Maran, A Tropsha - Molecular Informatics, 2016 - Wiley Online Library
Abstract Thousands of (Quantitative) Structure-Activity Relationships (Q) SAR models have been described in peer-reviewed publications; however, this way of sharing seldom makes models available for the use by the research community outside of the developer's
Cite Save More

ToxCast EPA in Vitro to in Vivo Challenge: Insight into the Rank-I Model
S Novotarskyi, A Abdelaziz, Y Sushko... - Chemical research in ..., 2016 - ACS Publications
The ToxCast EPA challenge was managed by TopCoder in Spring 2014. The goal of the challenge was to develop a model to predict the lowest effect level (LEL) concentration based on in vitro measurements and calculated in silico descriptors. This article summarizes
Cited by 3 Related articles All 3 versions Web of Science: 2 Cite Save More

Trust, But Verify II: A Practical Guide to Chemogenomics Data Curation

US Government Information

One stop source for US Government Information

HOME | CONSUMER | DEFENSE & INTERNATIONAL RELATIONS | EDUCATION & EMPLOYMENT | FAMILY, HOME, & COMMUNITY | HEALTH | MONEY | PUBLIC SAFETY & LAW | REFERENCE & SCIENCE & TECHNOLOGY | ABOUT

EDSP Prioritization: Collaborative Estrogen Receptor Activity Prediction Project (CERAPP) (SOT)

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

regulations.gov

Your Voice in Federal Decision-Making

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

Docket Folder Summary | View all documents and comments in this Docket

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

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

+ View More Docket Details

Primary Documents View All (2)

Meetings: Federal Insecticide, Fungicide, and Rodenticide Act Scientific Advisory Panel

ory Panel

EPA US Environmental Protection Agency

Español | 中文: 繁體版 | 中文: 简体版 | Tiếng Việt | 한국어

Learn the Issues | Science & Technology | Laws & Regulations | About EPA

Search EPA.gov

Related Topics: Safer Chemicals Research

Contact Us | Share

Safer Chemicals Research Update June 2016

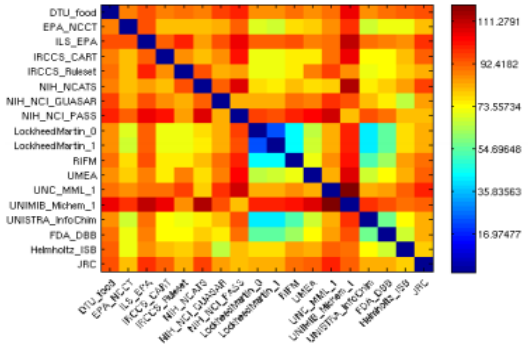
US EPA's Office of Research and Development provides quarterly updates, highlights, events and news about its chemical safety research. This is the June 2016 edition.

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

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

Consensus Modeling: Powering Prediction Through Collaboration

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



FEBRUARY 16-20 | AAAS 2017 ANNUAL MEETING | SERVING SOCIETY THROUGH SCIENCE POLICY | BOSTON

Adopting Alternative EDSP Assays

EDSP Tier 1 Battery of Assays	Model Alternative Development
Estrogen Receptor (ER) Binding ★	ER Model FY 2015
Estrogen Receptor Transactivation (ERTA) ★	ER Model FY 2015
Uterotrophic	ER Model FY 2015
Androgen Receptor (AR) Binding ★	AR Model FY 2016
Hershberger	AR Model FY 2016
Aromatase	STR Model FY 2016
Steroidogenesis (STR)	STR Model 2016
Female Rat Pubertal	ER, STR & THY Models FY 2017
Male Rat Pubertal	AR, STR & THY Models FY 2017
Fish Short Term Reproduction	ER, AR & STR Models FY 2017
Amphibian Metamorphosis	THY Model FY 2017

ER = estrogen receptor; AR = androgen receptor; STR = steroidogenesis; THY = thyroid

Slide with courtesy of Dr. N. Kleinstreuer

From CERAPP to CoMPARA : **Collaborative Modeling Project for Androgen Receptor Activity**

- Follow the steps of CERAPP
- Involve more research groups
- Increase the size of the prioritization set
- Use data from the combined ToxCast AR assays
- Collect and curate data from the literature for validation
- Use the previously designed workflows and code
- Use agonists, antagonists, and binding data
- Build continuous and classification models
- Adopt a similar approach for consensus modeling

CoMPARA participants: 34 international groups

CERAPP

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

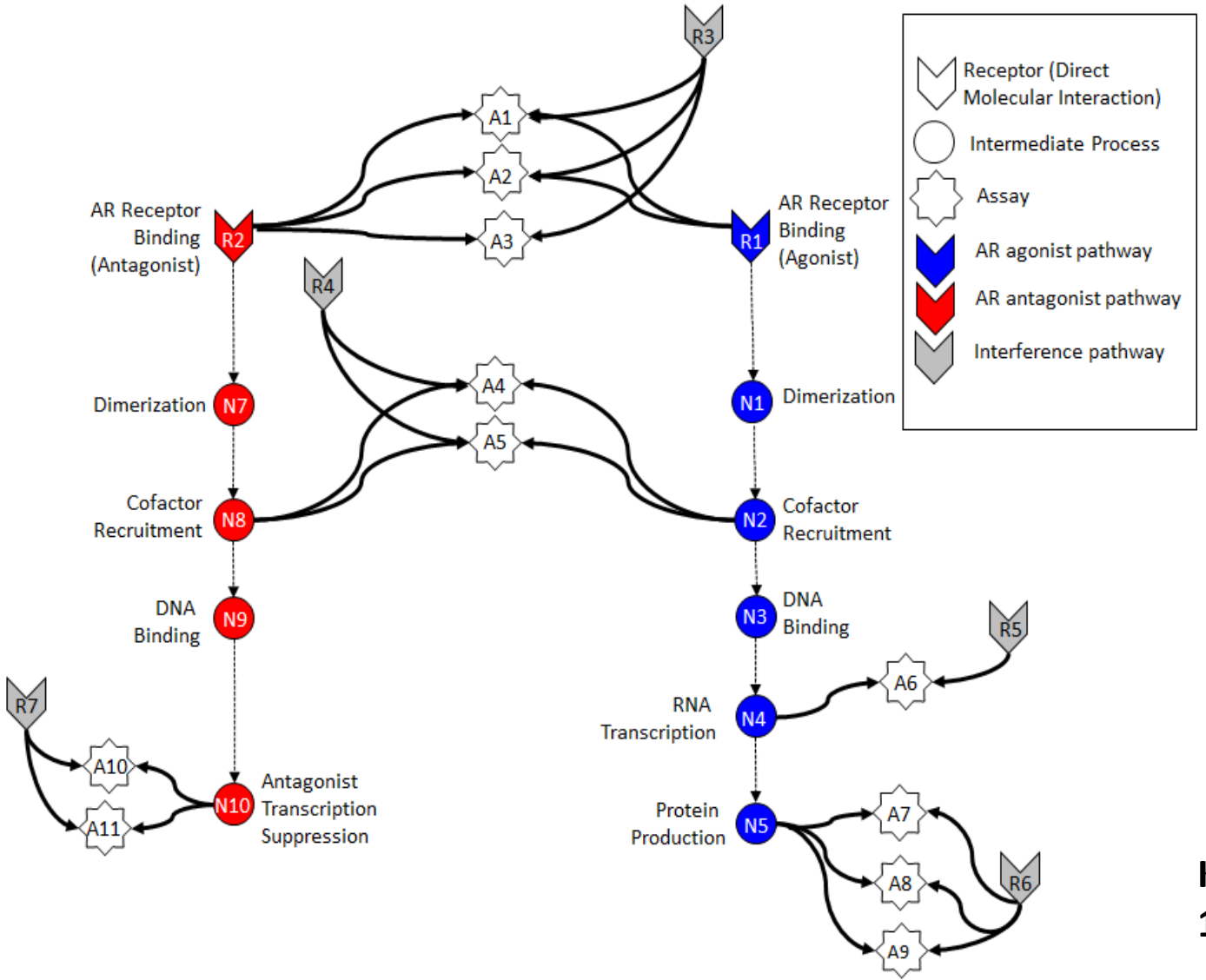
New groups

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

Plan of the project

1: Training and prioritization sets NCCT/ EPA	- ToxCast assays for training set data - AUC values and discrete classes for reg/class modeling - QSAR-ready training set and prioritization set
2: Experimental validation set NCCT/ EPA	- Collect and clean experimental data from the literature - Prepare validation sets for qualitative and quantitative models
3: Modeling & predictions All participants	- Train/refine the models based on the training set - Deliver predictions and applicability domains for evaluation
4: Model evaluation NCCT/ EPA	- Evaluate the predictions of each model separately - Assign a score for each model based on the evaluation step
5: Consensus predictions NCCT/ EPA	- Use the weighting scheme based on the scores to generate the consensus - Use the same validation set to evaluate consensus predictions
6: Manuscript writing All participants	- Descriptions of modeling approaches for each individual model - Input of the participants on the draft of the manuscript

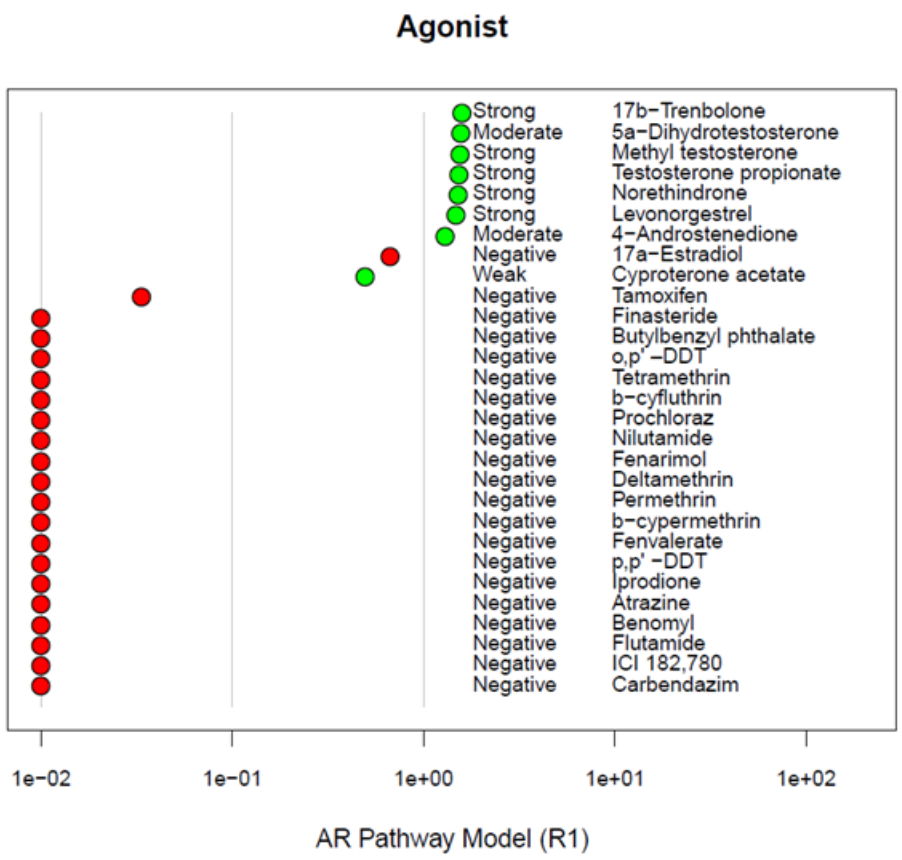
Tox21/ToxCast AR Pathway Model

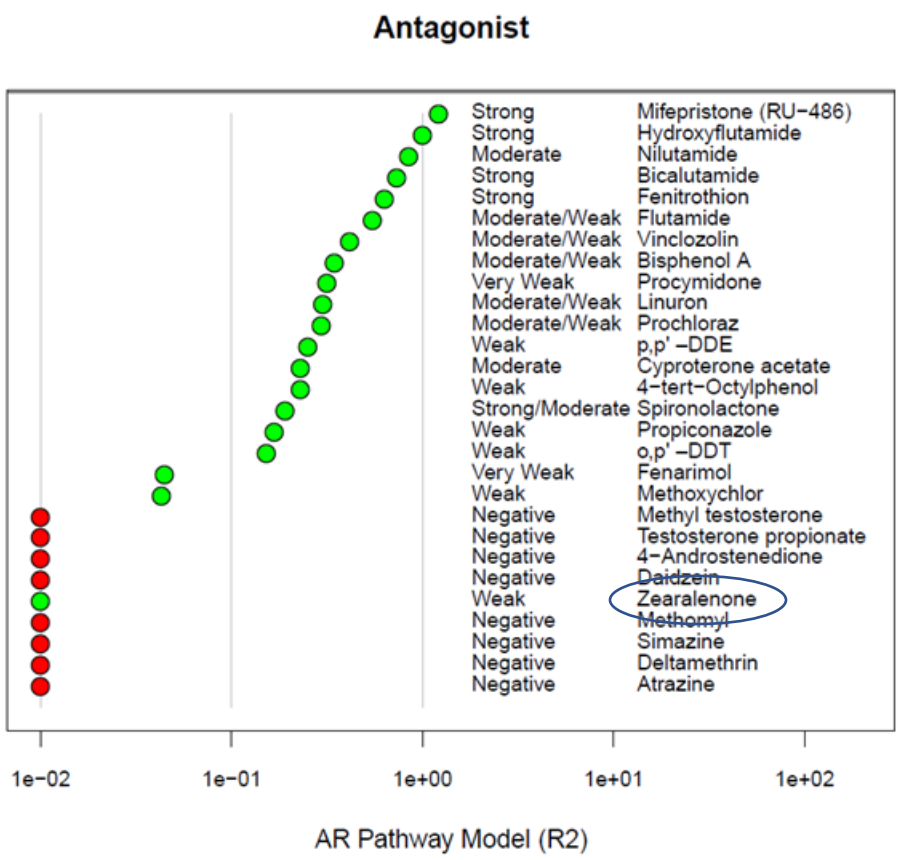


ToxCast High Throughput Screening AR assays		
Assay Name	**Biological Process**	**Assay #**
NVS_NR_hAR	receptor binding	1
NVS_NR_cAR	receptor binding	2
NVS_NR_rAR	receptor binding	3
OT_AR_ARSRC1_0480	cofactor recruitment	4
OT_AR_ARSRC1_0960	cofactor recruitment	5
ATG_AR_TRANS	mRNA induction	6
OT_AR_ARELUC_AG_1440	gene expression	7
Tox21_AR_BLA_Agonist_ratio	gene expression	8
Tox21_AR_LUC_MDAKB2_Agonist	gene expression	9
Tox21_AR_BLA_Antagonist_ratio	gene expression	10
Tox21_AR_LUC_MDAKB2_Antagonist	gene expression	11
Tox21_AR_LUC_MDAKB2_Antagonist*	gene expression	12

Kleinstreuer et al. (2016) Chem. Res. Toxicol. DOI: 10.1021/acs.chemrestox.6b00347

AR Pathway Model Performance

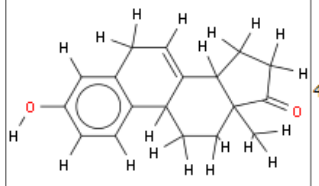
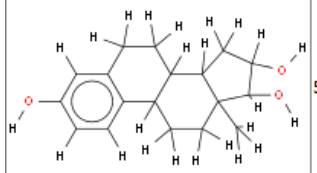
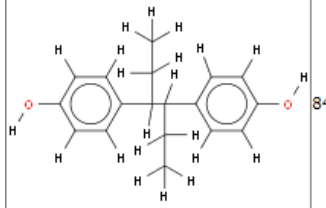
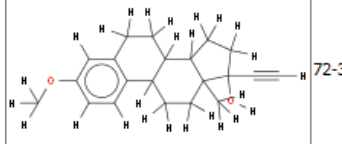




Kleinstreuer et al. (2016) Chem. Res. Toxicol. DOI: 10.1021/acs.chemrestox.6b00347

The one “false negative” was identified by confirmation assay results.

Training set: SDF file structure

Molecule	CASRN	Name	Canonical	InChI Code	InChI Key	Agonist	Antagonist	Binding	Agonist_Class	Antagonist_Class	Binding_Class
	474-86-2	Equilin	CC12CCC3c4ccc(O)c...	InChI=1S/C18H...	WKRLQDKEXYKH...	0.983	0.0	0.983	1	0	1
	50-27-1	Estriol	CC12CCC3C(CCc4cc...	InChI=1S/C18H...	PROQIPRRNZUX...	0.942	0.0	0.942	1	0	1
	84-16-2	meso-Hexes...	CCC(C(CC)c1ccc(O)...	InChI=1S/C18H...	PBBGSZCBWVPO...	0.879	0.0	0.879	1	0	1
	72-33-3	Mestranol	CC12CCC3C(CCc4cc...	InChI=1S/C21H...	IMSSROKUHAO...	0.858	0.0	0.858	1	0	1

1720 unique structures

Agonist: ~50 actives
Antagonist: ~160 actives
Binding: ~170 actives

false positives & false negatives excluded".

Prediction set

- CERAPP list: 32,464 unique QSAR-ready structures (organic, no mixtures…)
  - EDSP Universe (10K)
  - Chemicals with known use (40K) (CPCat & ACToR)
  - Canadian Domestic Substances List (DSL) (23K)
  - EPA DSSTox – structures of EPA/FDA interest (15K)
  - ToxCast and Tox21 (In vitro ER data) (8K)

  - CERAPP-DSSTox registered 29,904 QSAR ready => 45,981 GSIDs

- EINECS: European INventory of Existing Commercial chemical Substances
  - ~60k structures
  - ~55k QSAR-ready structures
  - ~38k non overlapping with the CERAPP list
  - ~18k overlap with DSSTox

29,904 + 17984 = 47,888 QSAR ready structures (with DSSTox GSIDs!)

SDF file contains 2D standardized QSAR-ready structure + GSID

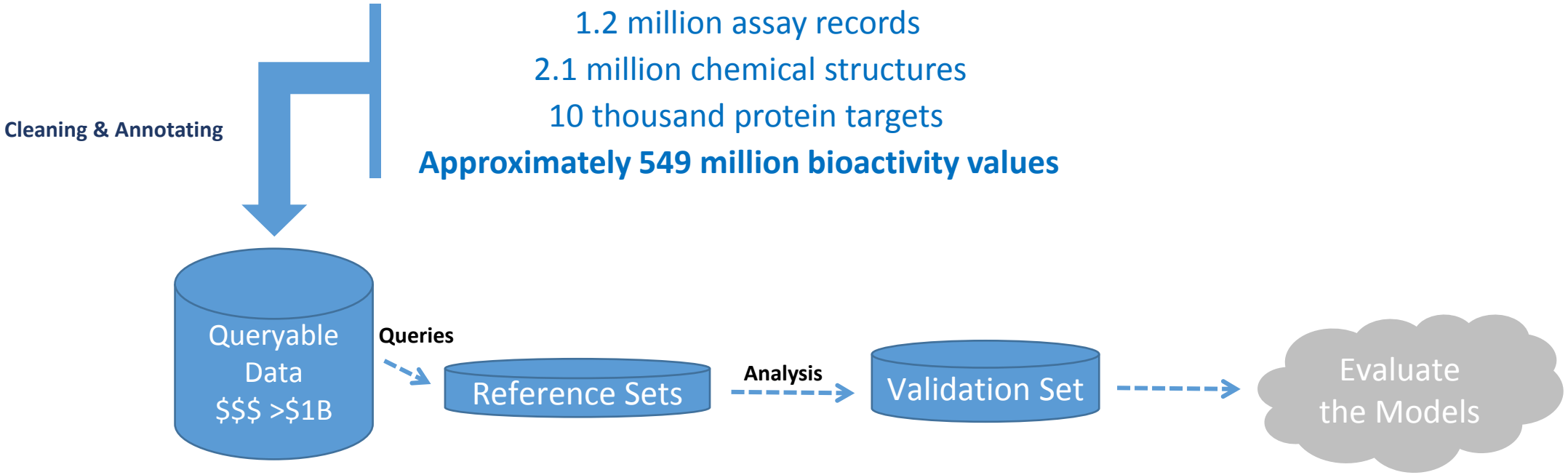
Validation set

Scrub Chem

PubChem

ChEMBL Toxcast Tox21
BindingDB MLSP etc.

1.2 million assay records
2.1 million chemical structures
10 thousand protein targets
Approximately 549 million bioactivity values

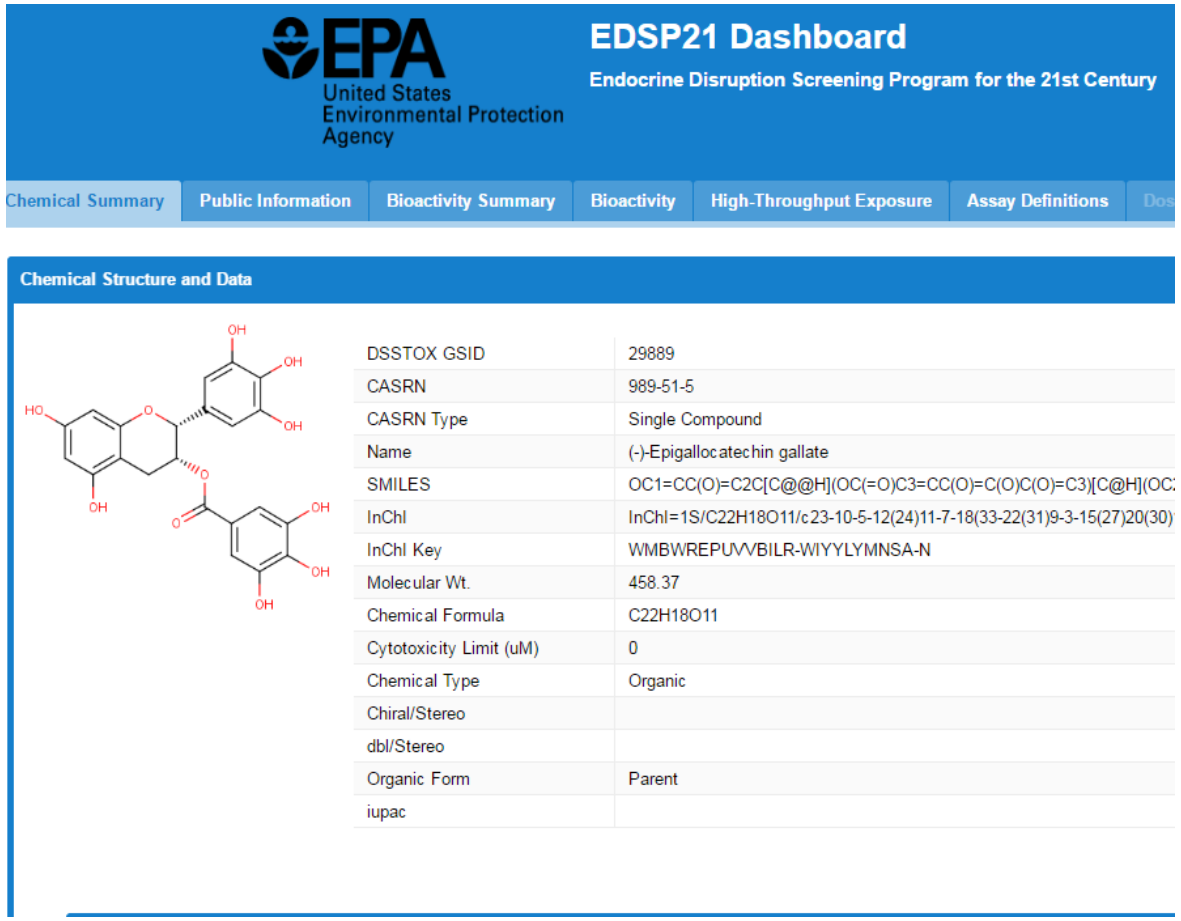


Slide with courtesy of Dr. J. Harris

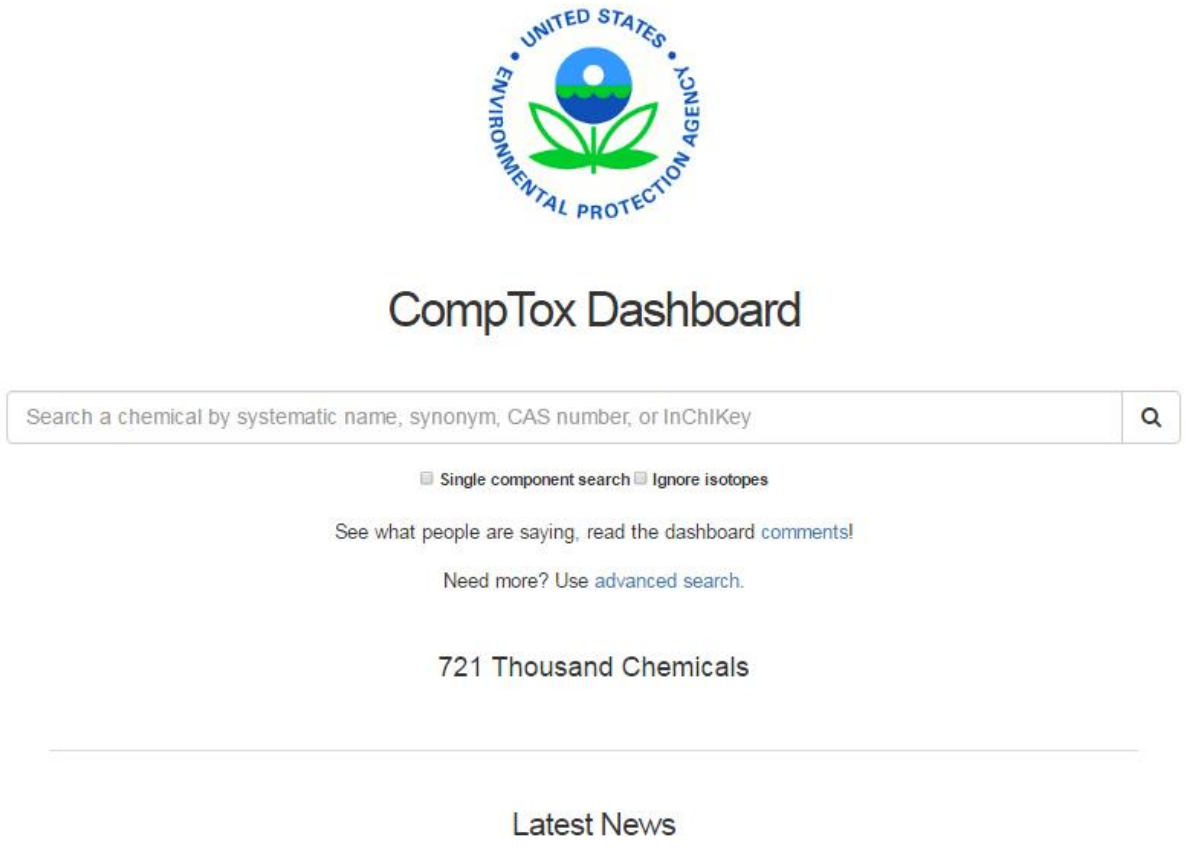
Harris and Judson (in preparation)

Online Publication of results

EDSP dashboard: http://actor.epa.gov/edsp21/



ICD dashboard: https://comptox.epa.gov/dashboard/



Summary

- Prioritized tens of thousands of chemicals for ER & AR in a fast accurate and economic way to help with the EDSP program.
- Generated high quality data and models that can be reused
- Free & open-source code and workflows
- Published manuscripts in peer reviewed journals
- Data and predictions available for visualization on the EDSP dashboard: http://actor.epa.gov/edsp21/

Acknowledgements

National Center for Computational Toxicology, US EPA

CERAPP participants

CoMPARA participants

Thank you for your attention

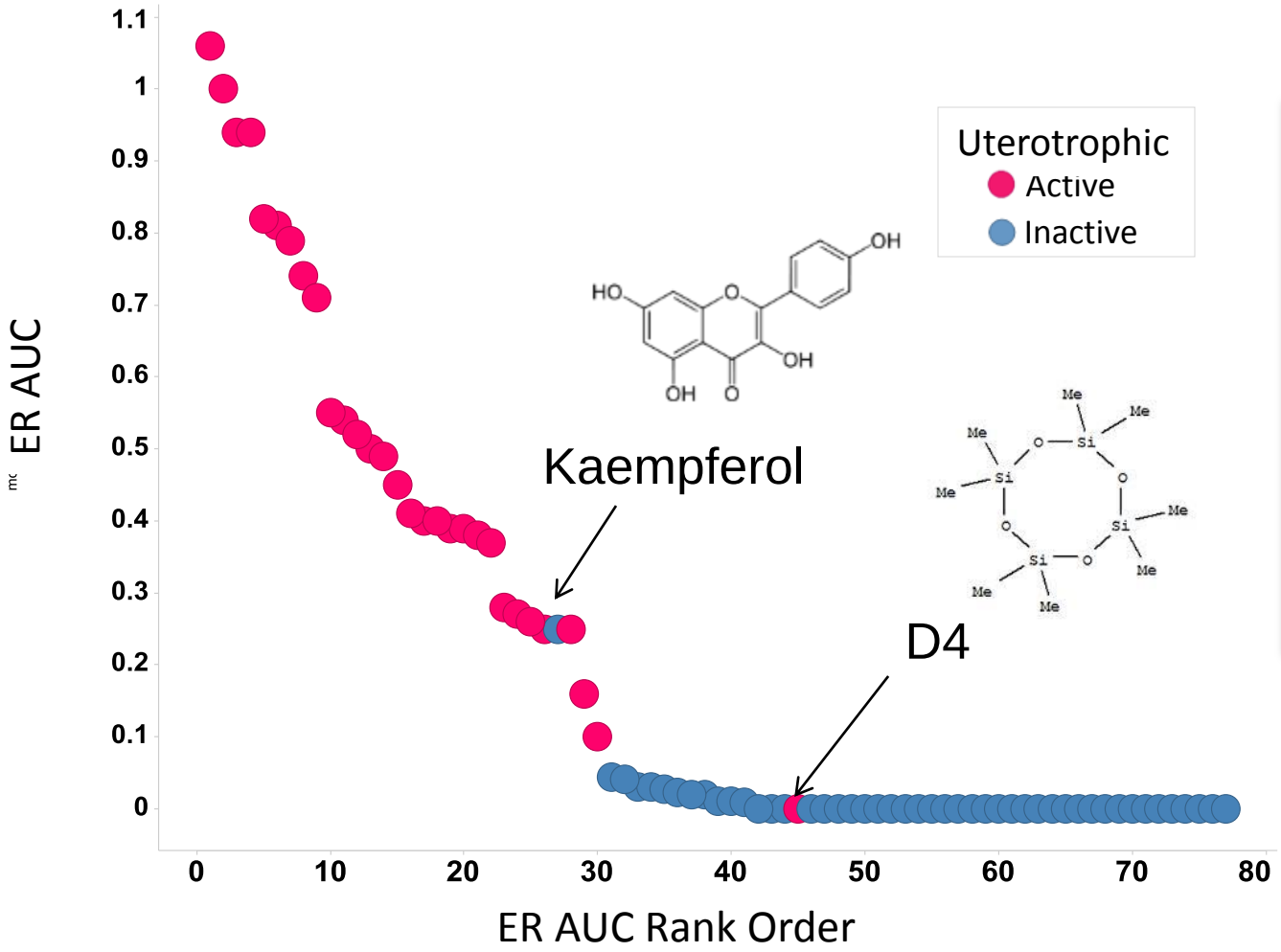
Question

OR

Comment

ER Model Performance

In Vivo Reference Chemicals



True Positive	29
True Negative	46
False Positive	1
False Negative	1
Accuracy	**0.97**
Sensitivity	**0.97**
Specificity	**0.97**

Browne et al. 2015 ES&T