Consensus Models to Predict Endocrine Disruption for All Human-Exposure Chemicals

Kamel Mansouri
Richard Judson

National Center for Computational Toxicology, U.S. EPA, RTP, NC, USA

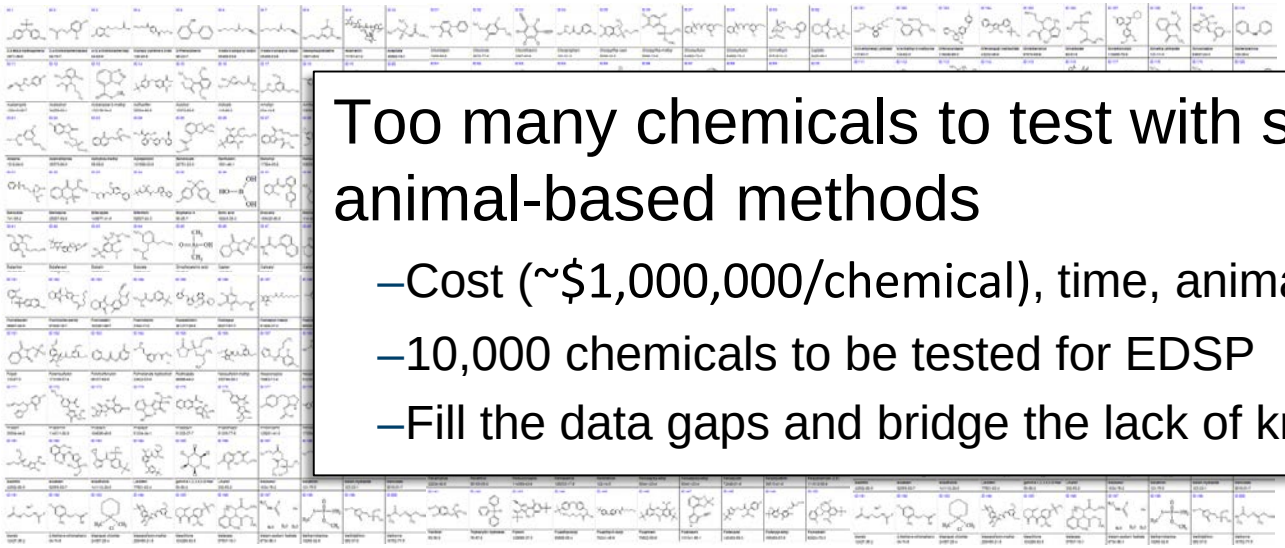
AAAS annual meeting
18 February 2017

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

Background and Goals

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

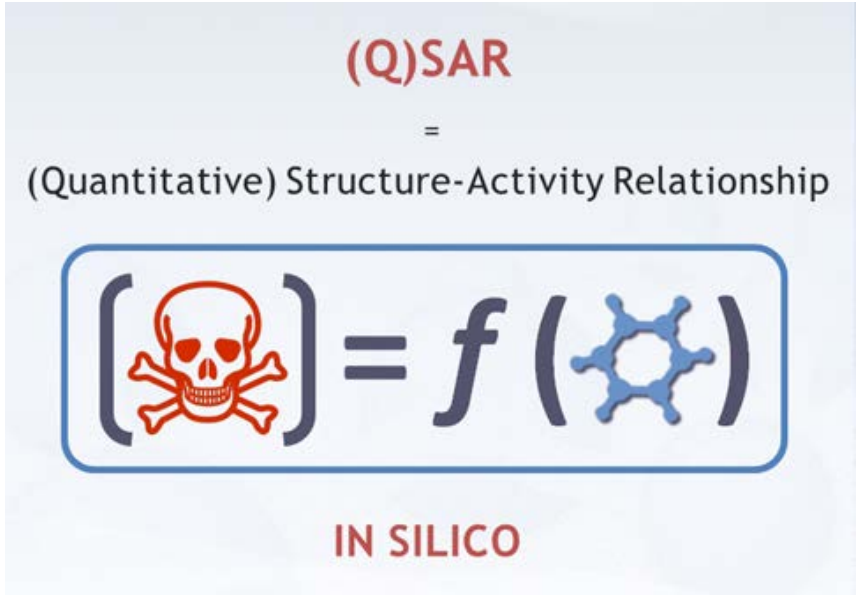
Problem Statement



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)

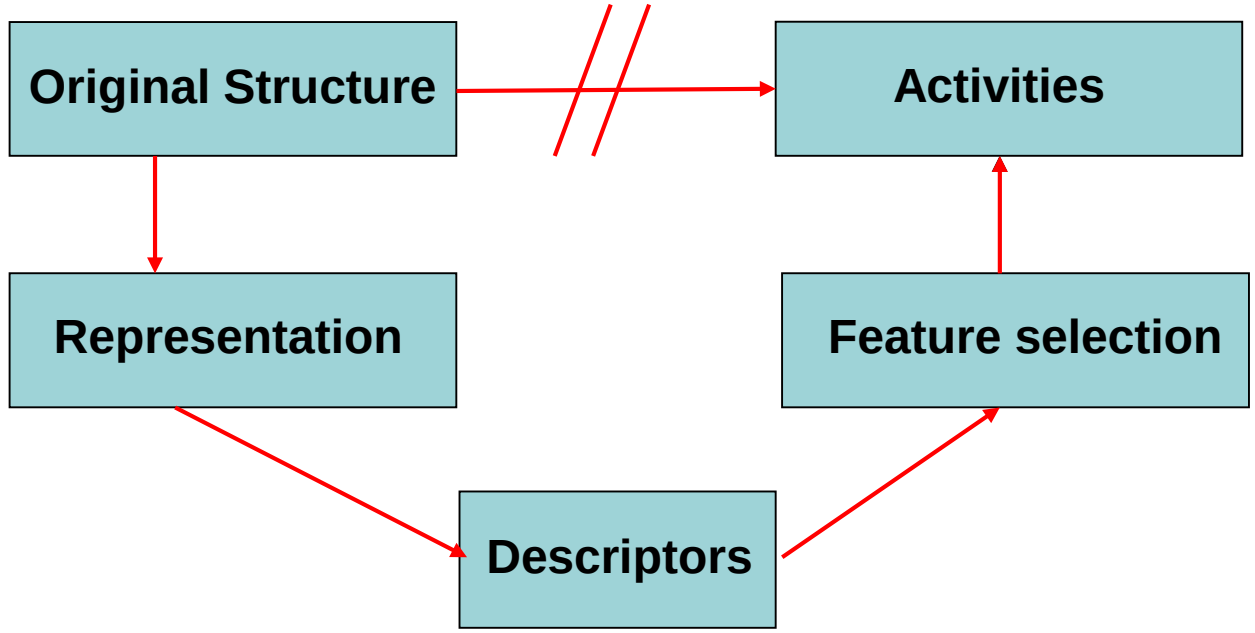
Congenericity principle: QSARs correlate, within congeneric series of compounds, their chemical or biological activities, either with certain structural features or with atomic, group or molecular descriptors.

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

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

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

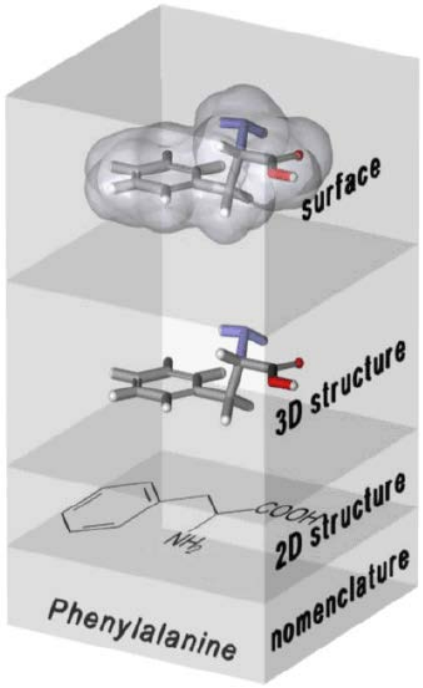
b_i - ***fitted parameters***

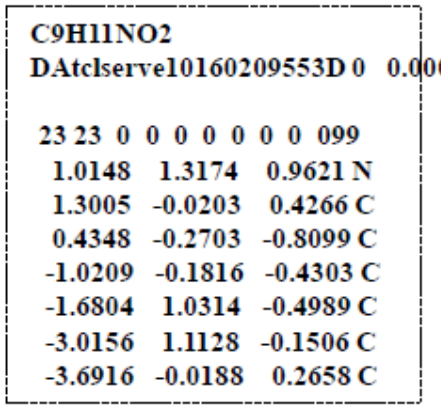


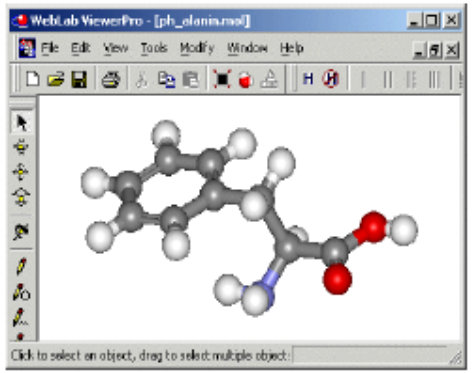
Development of a QSAR model

- Curation of experimental data
- Preparation of training and test sets
  - Data may be noisy: limits prediction accuracy
- 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

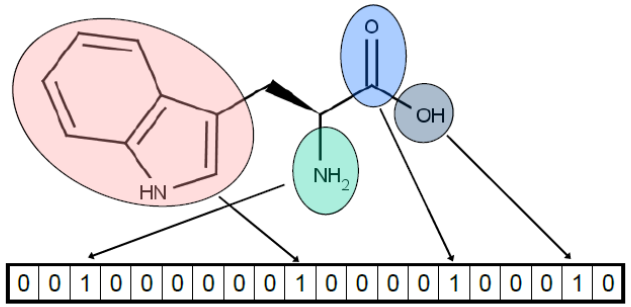
Molecular structures in the computer







Bitstrings in databases

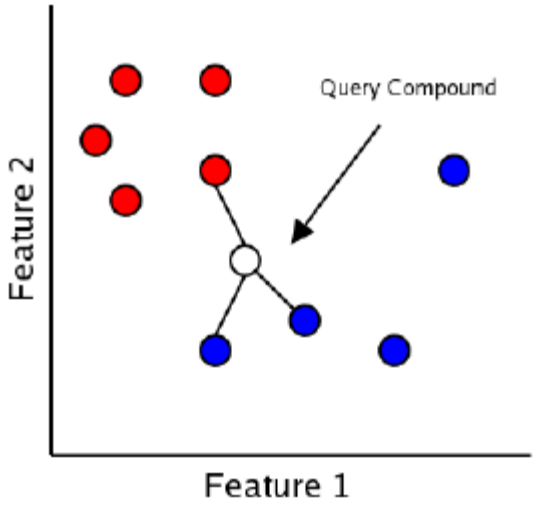


Fragmental keys & fingerprints

- substructural search
- read-across
- similarity search

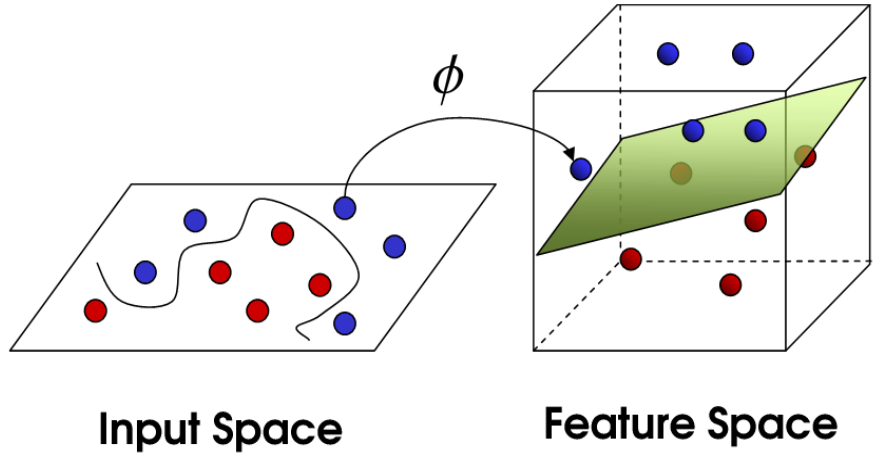
Classification methods

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



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

- **SVM: Support Vector Machines**



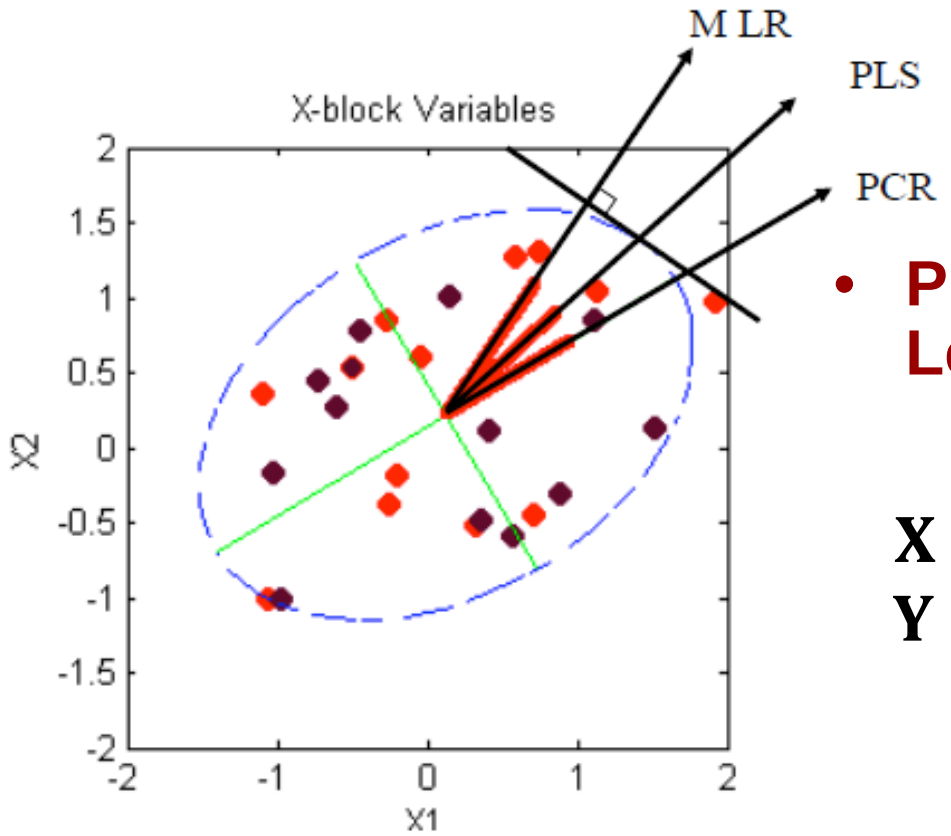
Kernel function maximizing the margin between the classes

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**

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

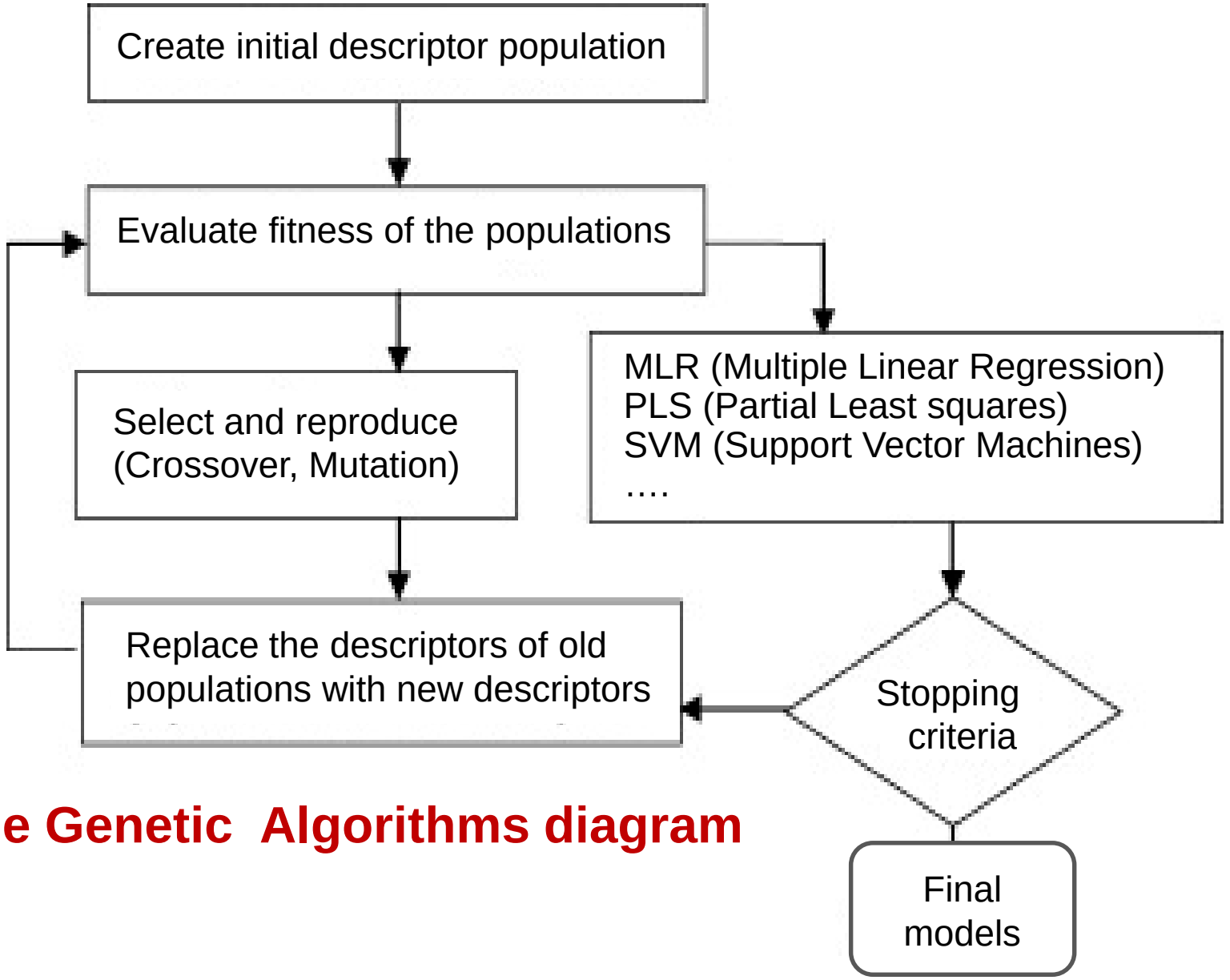
PLS is the vector on the PCR ellipse upon which MLR has the longest projection

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

Variable selection procedure

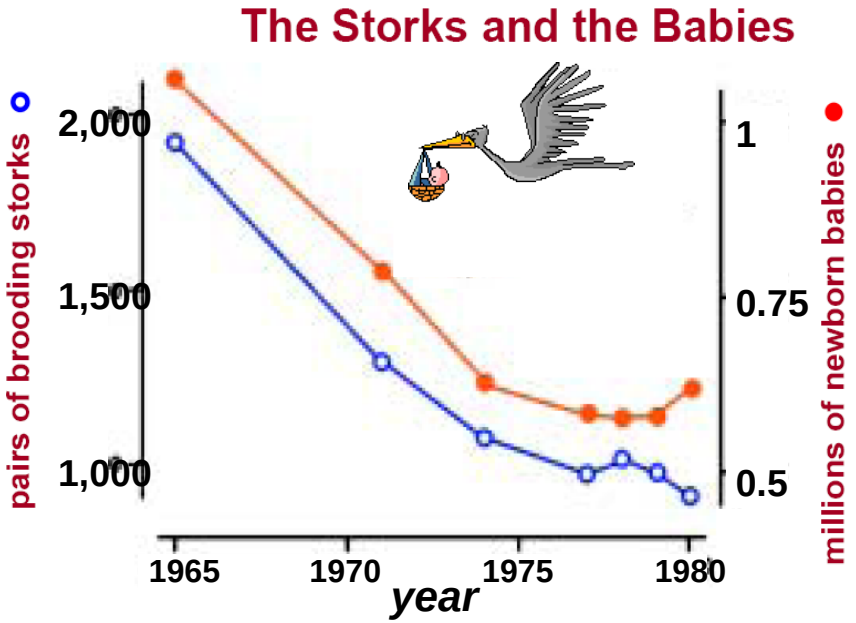
- **Many more descriptors than chemicals**
- **Many irrelevant descriptors**

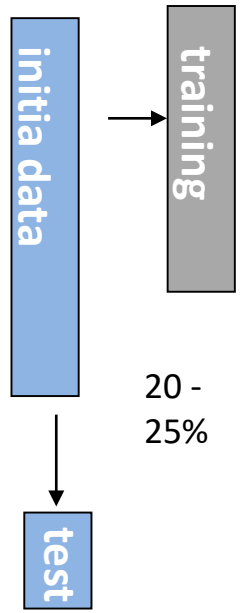
Only the most important descriptors are selected

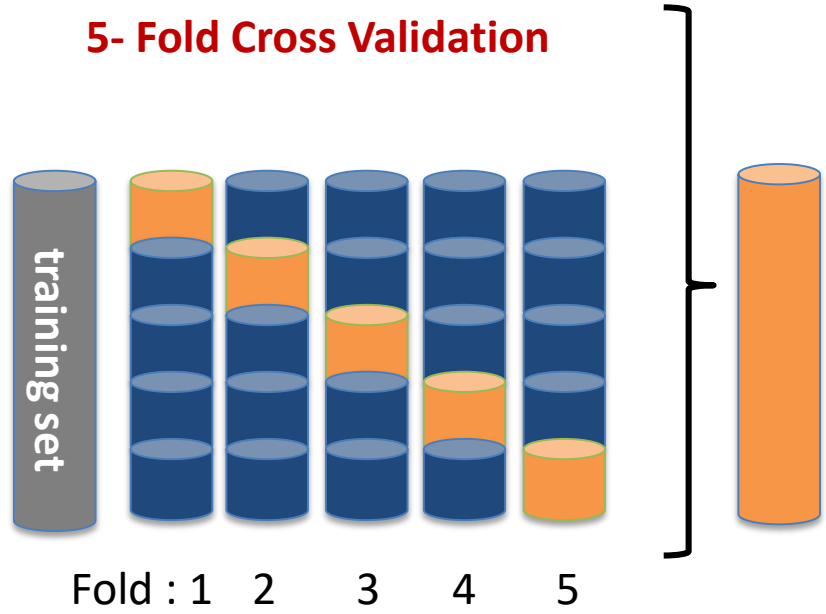


The Genetic Algorithms diagram

Validation by cross-validation 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**".

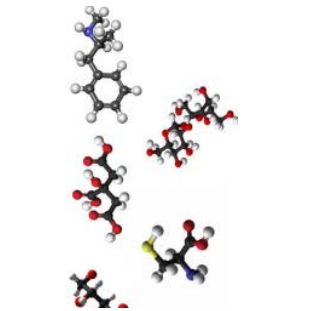
H. Sies, Nature 332, 495 (1988)

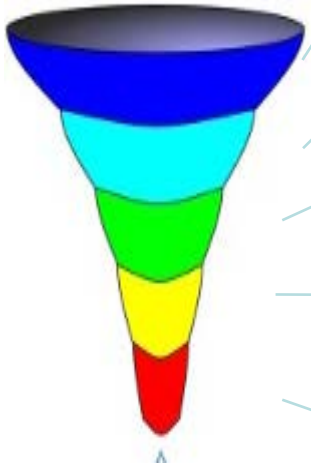
CERRAP : Collaborative Estrogen Receptor Activity Prediction Project

40 scientists, 17 research 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**

Structure curation procedure



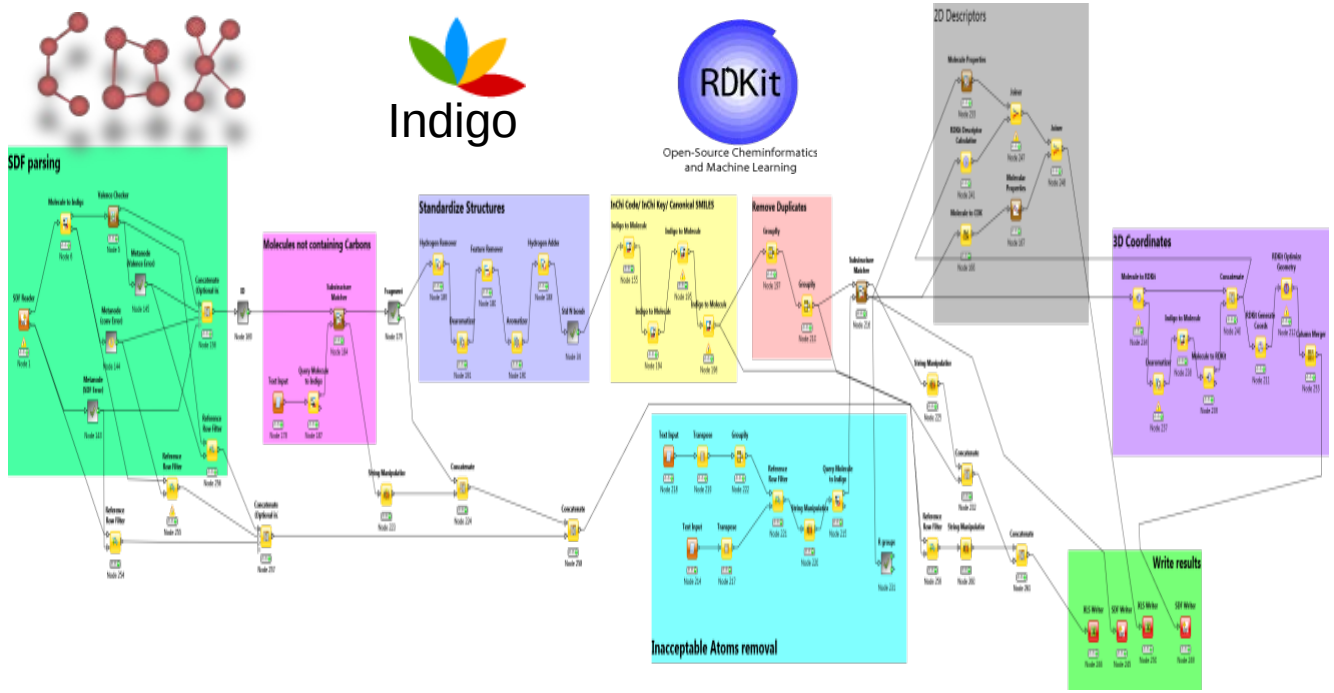


QSAR-ready structures

- Remove inorganics and mixtures
- Clean salts and counterions
- Normalize of tautomers
- Remove of duplicates
- Final inspection

Aim of the workflow:

- Combine different procedures and ideas
- Minimize the differences between the structures used for prediction
- Produce a flexible free and open source workflow to be shared



KNIME workflow

CERAPP data and results

Datasets of the project

- Training set: 1,677 chemicals **(ToxCast data)**
- Prediction set: 32,464 chemicals **(The Human Exposure Universe)**
- Evaluation set: 7,000 chemicals **(Literature: Tox21, FDA, METI…)**

Models received:

- Classification / Qualitative:
  - Binding: **22 models**
  - Agonists: **11 models**
  - Antagonists: **9 models**

Regression / Quantitative:

Binding: **3 models**
Agonists: **3 models**
Antagonists: **2 models**

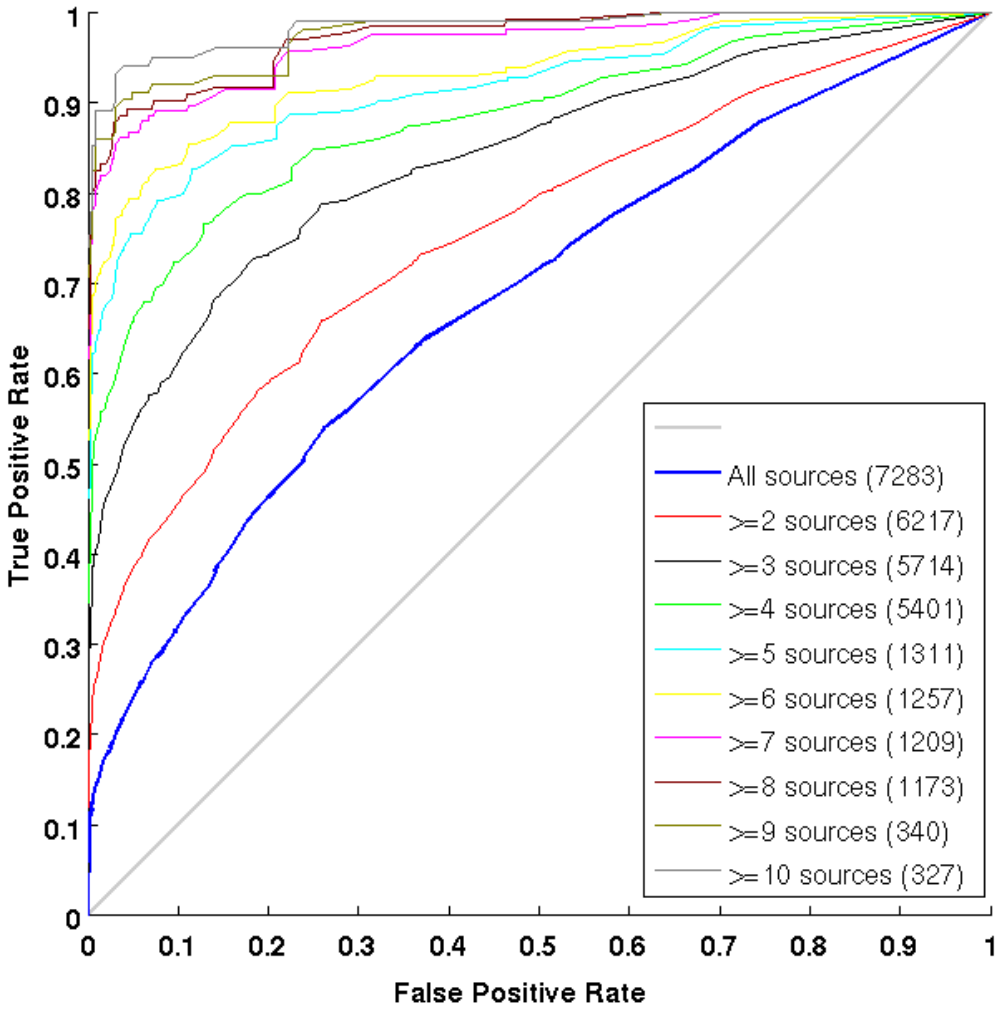
Consensus Qualitative Accuracy

Prediction Accuracy Strongly Depends on Data Quality

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

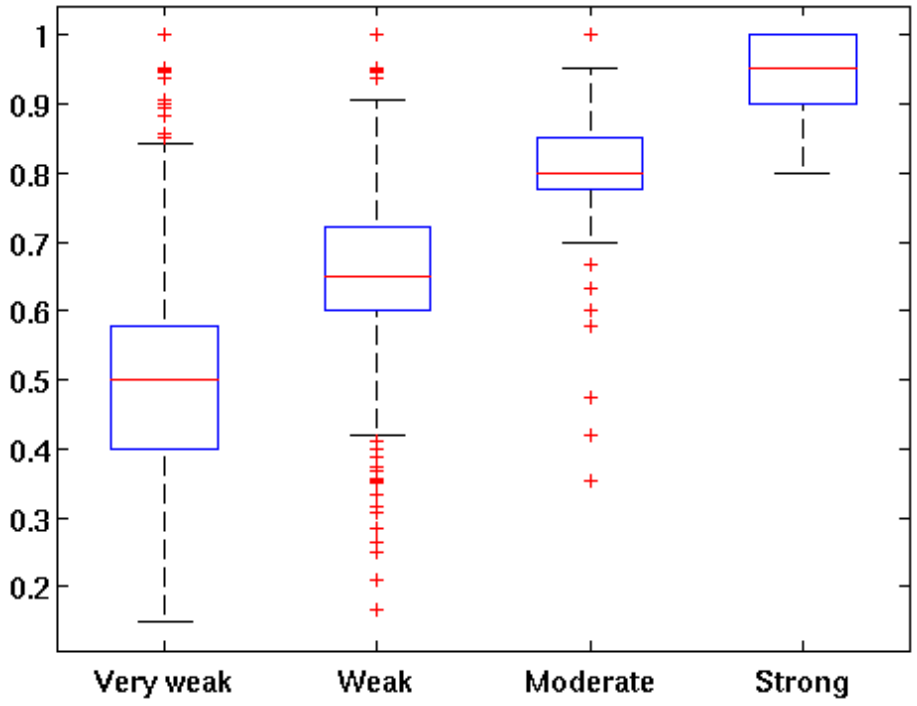
	ToxCast data		Literature data	
Observed\Predicted	Actives	Inactives	Actives	Inactives
Actives	83	6	597	1385
Inactives	40	1400	463	4838

	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

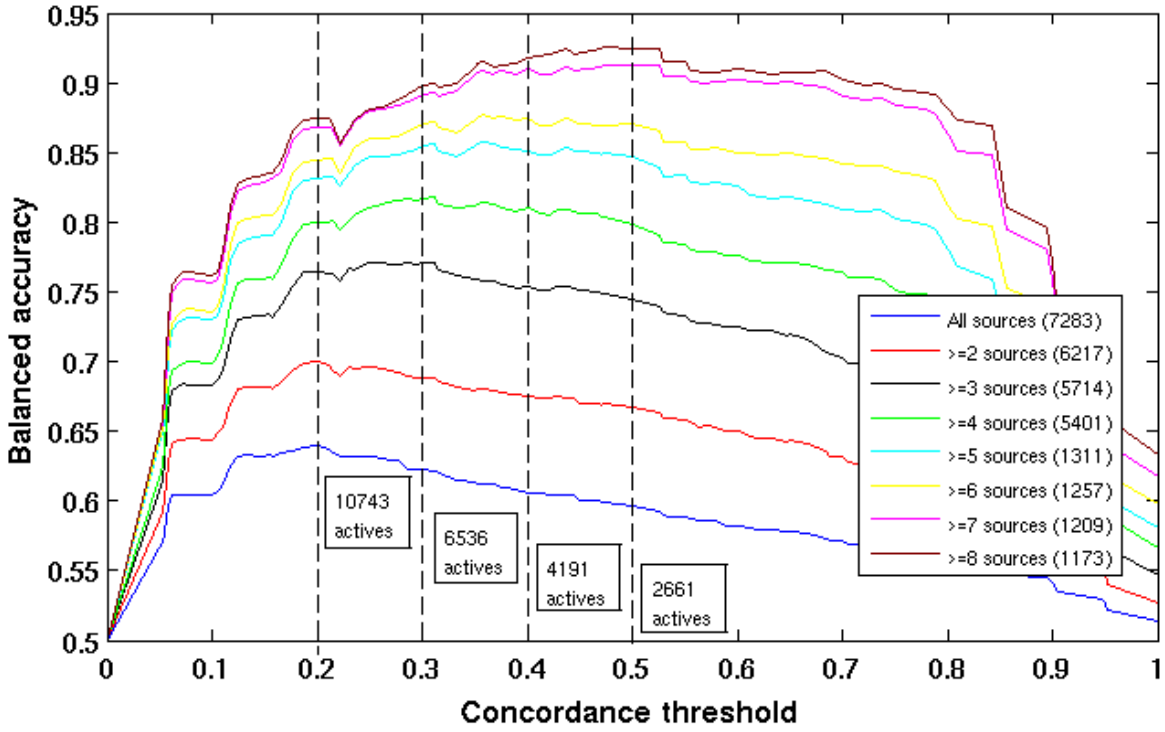


Consensus Quantitative Accuracy

- 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

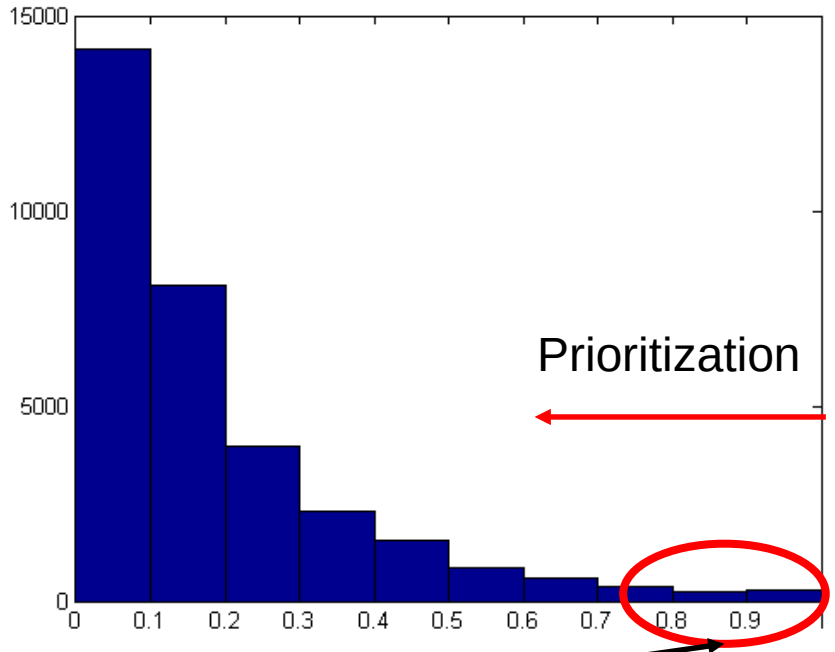


Box plot of the positive classes of the consensus model.



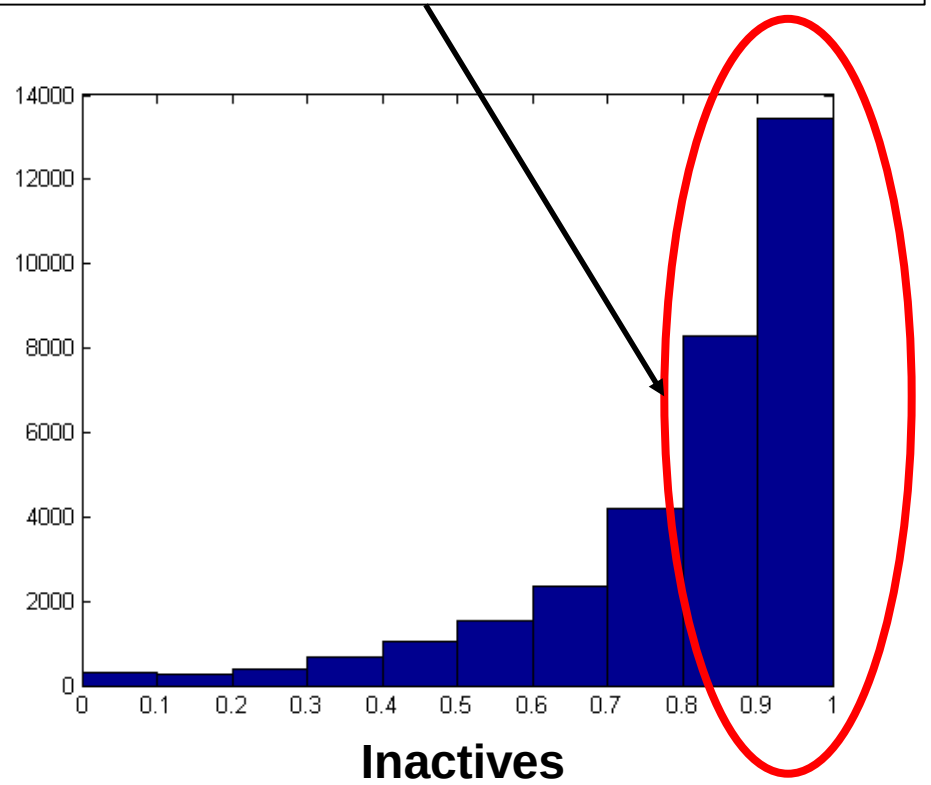
Variation of the balanced accuracy with positive concordance thresholds

Concordance of the qualitative models



Prioritization

Most models predict most chemicals as inactive



Only 757 chemicals have >75% positive concordance

⇨ Only a small fraction of chemicals require further testing!

ehp ENVIRONMENTAL HEALTH PERSPECTIVES

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

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
environment. It is well known that some environmental chemicals n

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

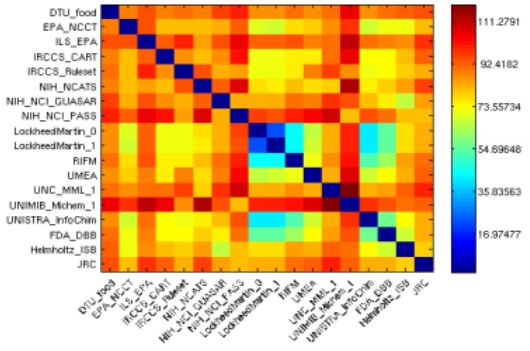
US EPA's Office of Research and Development provides quarterly updates, highlights, events and news about its chemical safety research. This is the June 2016 edition.

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 BOSTON
SERVING SOCIETY THROUGH SCIENCE POLICY

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

- Follow the CERAPP framework
- Use larger size prioritization set
- Use data from the combined ToxCast AR assays
- Collect and curate data from the literature for validation
- Use agonists, antagonists, and binding data
- Build continuous and classification models
- Similar approach for consensus modeling

CoMPARA participants: 34 international groups

United States Environmental Protection Agency

From 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 research 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**

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/

United States Environmental Protection Agency

Thank you for your attention