Virtual screening of chemicals for endocrine disrupting activity through CERAPP and CoMPARA projects

Kamel Mansouri[1], Nicole Kleinstreuer[2], Chris Grulke[3], Ann Richard[3], Imran Shah[3], Antony J. Williams[3] and Richard S. Judson[3]

1 ScitoVation LLC, RTP, NC, USA; **2** NTP Interagency Center for the Evaluation of Alternative Toxicological Methods, RTP, NC, USA; **3** National Center for Computational Toxicology, U.S. EPA, RTP, NC, USA

Introduction

Endocrine disrupting chemicals (EDCs) are xenobiotics that mimic the interaction of natural hormones at the receptor level and alter synthesis, transport and metabolism pathways. The prospect of EDCs causing adverse health effects in humans and wildlife has led to the development of scientific and regulatory approaches for evaluating bioactivity. This need is being partially addressed by the use of high-throughput screening (HTS) in vitro approaches and computational modeling. In the framework of the Endocrine Disruptor Screening Program (EDSP), the U.S. EPA led two worldwide consortiums to "virtually" (i.e., in silico) screen chemicals for their potential estrogenic and androgenic activities. The Collaborative Estrogen Receptor (ER) Activity Prediction Project (CERAPP) [1] predicted activities for 32,464 chemicals and the Collaborative Modeling Project for Androgen Receptor (AR) Activity (CoMPARA) generated predictions on the CERAPP list with additional simulated metabolites, totaling 55,450 unique structures. Modelers and computational toxicology scientists from 30 international groups contributed structure-based models and results for activity prediction to one or both projects, with methods ranging from QSARs to docking to predict binding, agonism and antagonism activities. Models were based on a common training set of 1746 chemicals having ToxCast/Tox21 HTS in vitro assay results (18 assays for ER and 11 for AR) integrated into computational networks. The models were then validated using curated literature data from different sources (~7,000 results for ER and ~5,000 results for AR). To overcome the limitations of single approaches, CERAPP and CoMPARA models were each combined into consensus models reaching high predictive accuracy. These consensus models were extended beyond the initially designed datasets by implementing them into the free and open-source application OPERA to avoid running every single model on new chemicals [2]. This implementation was used to screen the entire EPA DSSTox database of ~750,000 chemicals and predicted ER and AR activity is made freely available on the CompTox Chemistry dashboard (https://comptox.epa.gov/dashboard) [3].

Participants

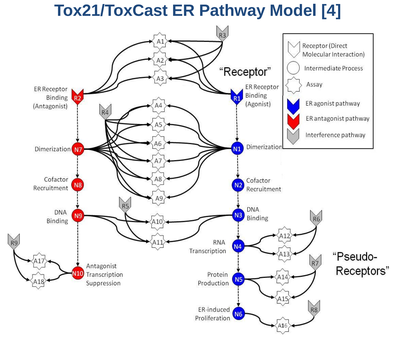
Group ID	Institution	Country	CERAPP	CoMPARA
ATSDR	Agency for Toxic Substances and Disease Registry. CDC.	USA		✓
DTU	Technical University of Denmark.	Denmark	✓	✓
ECUST	East China University of Science and Technology.	China		✓
IBMC	Institute of Biomedical Chemistry.	Russia		✓
IDEA	IdeaConsult, Ltd.	Bulgaria		✓
ILS	Integrated Laboratory Systems, Inc.	USA	✓	
INSLA	University of Insubria.	Italy		✓
INSLA	Lanzhou University.	China		
IRFMN	Istituto di Ricerche Farmacologiche "Mario Negri".	Italy	✓	✓
JRC	Joint Research Centre of the European Commission.	Italy	✓	
LM	Lockheed Martin IS&GS.	USA	✓	✓
MTI	Molecules Theurapetiques In silico.	France		✓
NCATS	National Center for Advancing Translational Sciences.	USA	✓	✓
NCCT	National Center for Computational Toxicology. EPA.	USA	✓	✓
NCI	National Cancer Institute.	USA	✓	
NCSU	North Carolina State University.	USA	✓	✓
NCTR	National Center for Toxicological Research. FDA.	USA	✓	✓
NICEATM	NTP Interagency Center for the Evaluation of Alternative Toxicological Methods.	USA	✓	✓
NRMRL	National Risk Management Research Laboratory. EPA.	USA		✓
RIFM	Research Institute for Fragrance Materials, Inc.	USA	✓	
SWETOX	Swedish toxicology research center.	Sweden		✓
TARTU	University of Tartu.	Estonia		✓
TUM	Technical University Munich.	Germany	✓	✓
UFG	Federal University of Golas.	Brazil		✓
UMEA	University of UMEA.	Sweden	✓	✓
UNC	University of North Carolina.	USA	✓	✓
UNIBA	University of Bari.	Italy	✓	✓
UNIMIB	University of Milano-Bicocca.	Italy	✓	✓
UNISTRA	University of Strasbourg.	France	✓	✓
VCCLAB	Virtual Computational Chemistry Laboratory.	Germany	✓	✓

Project planning

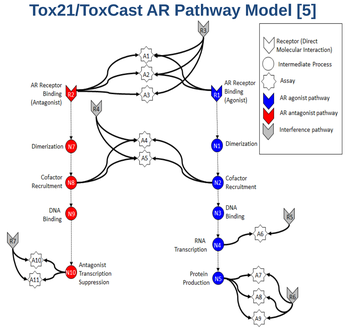
Steps	Tasks
1: Training and prioritization sets **Organizers**	- ToxCast assays for training set data - AUC values and discrete classes for continuous/classification modeling - QSAR-ready training set and prioritization set
2: Experimental validation set **Organizers**	- Collect and clean experimental 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 **Organizers**	- Evaluate the predictions of each model separately - Assign a score for each model based on the evaluation step
5: Consensus modeling **Organizers**	- 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

METHODS AND RESULTS

Tox21/ToxCast ER Pathway Model [4]



Tox21/ToxCast AR Pathway Model [5]



The lists of chemicals prioritized for ER and AR activity:

- **CERAPP list: 32,464 unique QSAR-ready structures** (standardized, organic, no mixtures…)
  - EDSP Universe (10k)
  - Chemicals with known use (40k) (CPCat & ACToR)
  - Canadian Domestic Substances List (DSL) (23k)
  - EPA DSSTox (version 1)– structures of EPA/FDA interest (15k)
  - ToxCast and Tox21 (In vitro ER data) (8k)
- **CoMPARA: 55,450 unique QSAR-ready structures**
  - CERAPP list (32k)
  - EINECS European Inventory (60k)
  - ToxCast metabolites (6k)

	ER training data			AR training data		
	Active	Inactive	Total	Active	Inactive	Total
Binding	237	1440	1677	198	1464	1662
Agonist	219	1458	1677	43	1616	1659
Antagonist	41	1636	1677	159	1366	1525
Total	237	1440	1677	198	1648	1688

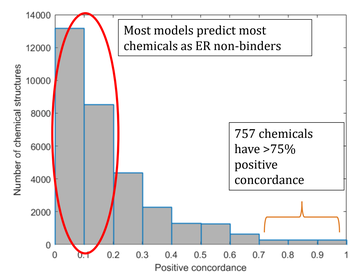
	ER validation data			AR validation data		
	Active	Inactive	Total	Active	Inactive	Total
Binding	1982	5301	7283	453	3429	3882
Agonist	350	5969	6319	167	4672	4839
Antagonist	284	6255	6539	355	3685	4040
Total	2017	7024	7522	487	4928	5273

	CERAPP participants models			CoMPARA participants models		
	Categorical	Continuous	Total	Categorical	Continuous	Total
Binding	21	3	24	32	5	37
Agonist	11	3	14	20	5	25
Antagonist	8	2	10	22	3	25
Total	40	8	48	74	13	87

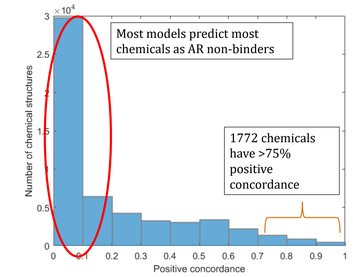
Categorical models	CERAPP validation scores			CoMPARA validation scores		
Group ID	**Binding**	**Agonist**	**Antagonist**	**Binding**	**Agonist**	**Antagonist**
ATSDR_IRFMN_1				0.69		
ATSDR_IRFMN_2				0.8		
ATSDR_IRFMN_3				0.85		
DTU_1	0.80	0.83		0.85	0.90	0.80
DTU_2		0.87				
ECUST				0.67		
IBMC_1				0.83	0.88	0.79
IBMC_2				0.86	0.87	0.86
IDEA				0.82		
INSLA				0.82	0.88	0.80
IRFMN_1	0.77			0.70		
IRFMN_2	0.77					
JRC	0.74	0.81	0.66			
LM_1	0.75	0.83	0.70	0.86	0.91	0.84
LM_2	0.70	0.77	0.62			
MTI				0.69	0.69	0.73
NCATS	0.67			0.66	0.79	0.58
NCCT_1	0.78			0.73		0.72
NCCT_2				0.85		0.84
NCCT_3				0.57		
NCI_1	0.84	0.91	0.77			
NCI_2	0.76	0.83	0.64			
NCSU				0.75	0.86	0.70
NCTR_1	0.84	0.89	0.78	0.84	0.89	0.82
NCTR_2	0.66					
ILS_NICEATM	0.79					
NRMRL_1				0.82	0.86	0.83
NRMRL_2				0.79	0.88	0.80
RIFM	0.69					
SWETOX_1				0.85	0.88	0.85
SWETOX_2				0.86	0.88	0.81
TARTU_1				0.85	0.88	0.78
TARTU_2				0.84	0.89	0.82
TUM				0.66	0.88	0.75
UFG				0.73	0.89	0.76
UMEA	0.76			0.88	0.92	0.77
UNC	0.73			0.82		
UNIBA	0.80	0.84	0.78	0.68		
UNIMIB_1	0.68			0.51		
UNIMIB_2	0.85					
UNISTRA	0.73	0.79		0.77	0.88	0.67
VCCLAB	0.80	0.87	0.69	0.86	0.88	0.76

The score formula: $S = 0.3 * (goodness\ of\ fit) + 0.45 * (predictivity) + 0.25 * (robustness)$. Where the goodness of fit (training set) and the predictivity (validation set) are functions of the balanced accuracy (BA) and the balance between the specificity (Sp) and Sensitivity (Sn): $0.7 * (BA) + 0.3 * (1 - |Sn - Sp|)$. Robustness of the model is the balance between training and validation set statistics: $1 - |BA_{Tr} - BA_{test}|$.

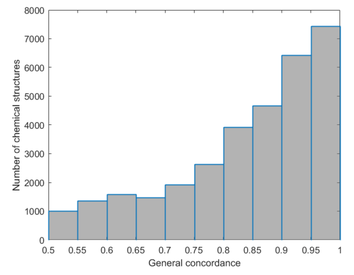
CERAPP models

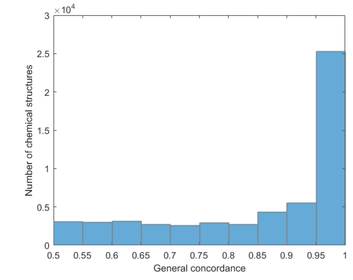


CoMPARA models

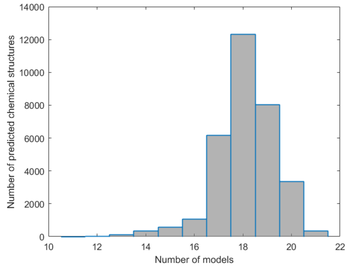


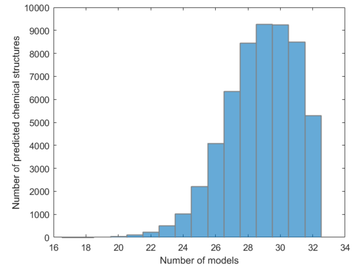
Concordance of binding models on the active compounds of the prediction sets.





High general concordance across all binding models included in the consensus.





Distributions of the number of the predicted chemical structures by all binding models.

Consensus models

CERAPP consensus validation

	Binding		Agonist		Antagonist	
	Training	Validation	Training	Validation	Training	Validation
Sn	0.93	0.58	0.85	0.94	0.67	0.18
Sp	0.97	0.92	0.98	0.94	0.94	0.90
BA	0.95	0.75	0.92	0.94	0.80	0.54

CoMPARA consensus validation

	Binding		Agonist		Antagonist	
	Training	Validation	Training	Validation	Training	Validation
Sn	0.99	0.69	0.95	0.74	1.00	0.61
Sp	0.91	0.87	0.98	0.97	0.95	0.87
BA	0.95	0.78	0.97	0.86	0.97	0.74

	CERAPP		CoMPARA	
	Active	Inactive	Active	Inactive
Binding	4001	28463	8202	40656
Agonist	2475	29989	1764	47094
Antagonist	2793	29671	9899	38959
Total	4001	28463	10623	47613

CoMPARA predictions

	ToxCast metabolites	
	Active	Inactive
Binding	1609	4983
Agonist	428	6164
Antagonist	1820	4772
Total	1989	6325

These consensus models predictions where model concordance is higher than 85% are being implemented into the free and open-source application OPERA in order to prioritize additional chemicals.

REFERENCES

1) Mansouri, K. et al. CERAPP: Collaborative estrogen receptor activity prediction project. *Environ. Health. Perspect.* **2016**; 124: 1023-1033.
2) Mansouri, K. et al. OPERA models for predicting physicochemical properties and environmental fate endpoints. Journal of Cheminformatics **2018** In Press, DOI: 10.1186/s13321-018-0263-1.
3) Williams, A. J. et al. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. Journal of Cheminformatics **2017**, 9, (1), 61.
4) Judson, R. et al. Integrated model of chemical perturbations of a biological pathway using 18 in vitro high-throughput screening assays for the estrogen receptor. *Toxicological Sciences.* **2015**; 148: 137-154.
5) Kleinstreuer N. C. et al. Development and Validation of a Computational Model for Androgen Receptor Activity. Chemical Research in Toxicology **2017** 30 (4), 946-964.

ACKNOWLEDGEMENTS

This work received funding from:
- Participation Program at the EPA/ORD, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and EPA.
- The Lush Prize 2017, category of young researchers (Americas), supporting animal free testing.

Disclaimer: The views expressed in this presentation are those of the authors and do not necessarily represent the U.S. EPA or NIEHS policies.