CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity

Mansouri, Kamel; Kleinstreuer, Nicole; Abdelaziz, Ahmed M.; Alberga, Domenico; Alves, Vinicius M.; Andersson, Patrik L.; Andrade, Carolina H.; Bai, Fang; Balabin, Ilya; Ballabio, Davide

Total number of authors: 72

Published in: Environmental Health Perspectives

Link to article, DOI: 10.1289/ehp5580

Publication date: 2020

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity

Kamel Mansouri,1,2,3 Nicole Kleinsteuere,4 Ahmed M. Abdelaziz,5 Domenico Alberga,6 Vinicius M. Alves,7,8 Patrik L. Andersson,9 Carolina H. Andrade,7 Fang Bai,10 Ilya Balabin,11 Davide Ballabio,12 Emilio Benfenati,14 Barun Bhattacharaj,15 Scott Boyer,16 Jingwen Chen,17 Viviana Cosson,12 Sherif Farag,6 Denis Fourches,8 Alfonso T. Garcia-Sosa,16 Paola Gramatica,15 Francesca Grisoni,12 Chris M. Gruke,1 Huixiao Hong,20 Dragos Horvath,21 Xin Hu,22 Ruili Huang,22 Nina Jeliazkova,23 Jiazhong Li,10 Xuehua Li,17 Huaxiang Liu,10 Serena Manganelli,14 Giuseppe F. Mangiafico,6,9 Uko Maran,19 Gilles Marcou,21 Todd Martin,22 Eugene Muratov,8 Duc-Trung Nguyen,2 Orazio Nicolotti,6 Nikolai G. Nikolov,13 Ulf Norinder,16 Ester Papa,15 Michel Petitjean,25 Geven Piir,19 Pavel Pogodin,16 Vladimir Porokov,20 Xianliang Qiao,17 Ann M. Richard,1 Alessandra Roncaglioni,14 Patricia Ruiz,27 Chetan Rupakheti,24,28 Sugunadevi Sakkiah,20 Alessandro Sangion,15 Karl-Werner Schramm,5 Chandrakshya Selvaraj,20 Imran Shah,1 Sulay Sild,19 Lixia Sun,29 Olivier Taboureau,25 Yun Tang,29 Igor V. Tetko,30,31 Roberto Todeschini,12 Weida Tong,20 Daniela Trisciuzzi,6 Alexander Tropsha,8 George Van Den Driessche,18 Alexandre Varnek,21 Zhongyu Wang,17 Eva B. Wedeby,13 Antony J. Williams,1 Hongbin Xie,17 Alexey V. Zakharov,22 Ziye Zheng,9 and Richard S. Judson1

1National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency (U.S. EPA), Research Triangle Park, North Carolina, USA; 2ScitoVation LLC, Research Triangle Park, North Carolina, USA; 3Integrated Laboratory Systems, Inc., Morrisville, North Carolina, USA; 4National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM), National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA; 5Technische Universität München, Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt, Department für Biowissenschaftliche Grundlagen, Weihenstephaner Stieg 23, 85350 Freising, Germany; 6Department of Pharmacy-Drug Design, University of Bari, Bari, Italy; 7Laboratory for Molecular Modeling and Drug Design, Faculty of Pharmacy, Federal University of Goiás, Goiânia, Brazil; 8Laboratory for Molecular Modeling, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA; 9Swedish Toxicology Sciences Research Center, Karolinska Institutet, Södertälje, Sweden; 10School of Pharmacy, Lanzhou University, China; 11Information Systems & Global Solutions (IS&GS), Lockheed Martin, USA; 12Milano Chemometrics and QSAR Research Group, Department of Earth and Environmental Sciences, University of Milano-Bicocca, Milan, Italy; 13Division of Risk Assessment and Nutrition, National Food Institute, Technical University of Denmark, Copenhagen, Denmark; 14Istituto di Ricerche Farmacologiche “Mario Negri”, IRCCS, Milan, Italy; 15QSAR Research Unit in Environmental Chemistry and Ecotoxicology, Department of Theoretical and Applied Sciences, University of Insubria, Varese, Italy; 16Swedish Toxicology Sciences Research Center, Karolinska Institutet, Södertälje, Sweden; 17School of Environmental Science and Technology, Dalian University of Technology, Dalian, China; 18Department of Chemistry, Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, USA; 19Institute of Chemistry, University of Tartu, Tartu, Estonia; 20Division of Bioinformatics and Biostatistics, National Center for Toxicology Research, U.S. Food and Drug Administration, Jefferson, Arkansas, USA; 21Laboratoire de Chimieinformatique—UMR7140, University of Strasbourg/CNRS, Strasbourg, France; 22National Center for Advancing Translational Sciences, National Institutes of Health, Rockville, Maryland, USA; 23IdaConsult, Ltd., Sofia, Bulgaria; 24National Risk Management Research Laboratory, U.S. EPA, Cincinnati, Ohio, USA; 25Computational Modeling of Protein-Ligand Interactions (CMPLI)–INSERM UMR 8251, INSERM ERL U1133, Functional and Adaptive Biology (FBA), Universite de Paris, Paris, France; 26Institute of Biomedical Chemistry IBMC, 10 Building 8, Pogodinskaya st., Moscow 119121, Russia; 27Computational Toxicology and Methods Development Laboratory, Division of Toxicology and Human Health Sciences, Agency for Toxic Substances and Disease Registry, Centers for Disease Control and Prevention, Atlanta, Georgia, USA; 28Department of Biochemistry and Molecular Biophysics, University of Chicago, Chicago, Illinois, USA; 29Department of Pharmaceutical Sciences, School of Pharmacy, East China University of Science and Technology, Shanghai, China; 30BIGCHEM GmbH, Neuberger, Germany; 31Helmholtz Zentrum München – German Research Center for Environmental Health (GmbH), Neuberger, Germany.

BACKGROUND: Endocrine disrupting chemicals (EDCs) are xenobiotics that mimic the interaction of natural hormones and alter synthesis, transport, or metabolic 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 addressed using high-throughput screening (HTS) in vitro approaches and computational modeling.

OBJECTIVES: In support of the Endocrine Disruptor Screening Program, the U.S. Environmental Protection Agency (EPA) led two worldwide consortia to virtually screen chemicals for their potential estrogenic and androgenic activities. Here, we describe the Collaborative Modeling Project for Androgen Receptor Activity (CoMPARA) efforts, which follows the steps of the Collaborative Estrogen Receptor Activity Prediction Project (CERAPP).
**METHODS:** The CoMPARA list of screened chemicals built on CERAPP’s list of 32,464 chemicals to include additional chemicals of interest, as well as simulated ToxCast™ metabolites, totaling 55,450 chemical structures. Computational toxicology scientists from 25 international groups contributed 91 predictive models for binding, agonist, and antagonist activity predictions. Models were underpinned by a common training set of 1,746 chemicals compiled from a combined data set of 11 ToxCast™/Tox21 HTS in vitro assays.

**RESULTS:** The resulting models were evaluated using curated literature data extracted from different sources. To overcome the limitations of single-model approaches, CoMPARA predictions were combined into consensus models that provided averaged predictive accuracy of approximately 80% for the evaluation set.

**DISCUSSION:** The strengths and limitations of the consensus predictions were discussed with example chemicals; then, the models were implemented into the free and open-source OPERA application to enable screening of new chemicals with a defined applicability domain and accuracy assessment. This implementation was used to screen the entire EPA DSSTox database of ~ 875,000 chemicals, and their predicted AR activities have been made available on the EPA CompTox Chemicals dashboard and National Toxicology Program’s Integrated Chemical Environment. https://doi.org/10.1289/EHP5580

**Introduction**

Humans are exposed to an increasingly high number of natural and synthetic chemical substances (Dionisio et al. 2015; Eggehy et al. 2012; Judson et al. 2009). These exogenous chemicals may have the potential to cause adverse health effects to humans and ecological species (Gray et al. 1997; Safe 1997). The endocrine system regulates a fragile hormonal equilibrium, which might be altered by chemicals that interfere with hormone signaling, e.g., by interacting with its different receptors. Over the last few decades, endocrine-disrupting chemicals (EDCs) have been linked to a large number of health issues, including neurological, developmental, reproductive, cardiovascular, metabolic, and immune system disorders (Colborn et al. 1993; Davis et al. 1993; Diamanti-Kandarakis et al. 2009; European Environmental Agency 2012; Martin et al. 2010; Skakkebaek et al. 2011; WHO 2013). The estrogen receptors (ER) and androgen receptors (AR) are among the most studied targets, with a variety of in silico (Bolger et al. 1998; Judson et al. 2015; Waller et al. 1996), in vitro (Chang et al. 2015; Fang et al. 2000; Rotroff et al. 2010; Shanle and Xu 2011; Soto et al. 1998), and in vivo (Kleinreuter et al. 2015; Mueller and Korach 2001; U.S. EPA 2011) EDC screening assays available.

The Endocrine Disruptor Screening Program (EDSP) of the U.S. Environmental Protection Agency (EPA) is one of the largest efforts to screen chemicals for endocrine-disrupting potential (U.S. EPA 2014b; U.S. EPA-OCSSP 2014, 2015). However, the time and cost to screen the approximately 10,000 chemicals required by EDSP through the entire battery of ToxCast™ endocrine disruptor assays, estimated at ~ $1 million USD per chemical, is untenable (HSIA 2009; U.S. EPA 2013, 2015). The EDSP has begun to address this resource issue by using in vitro high-throughput screening (HTS) assays included in the EPA’s ToxCast™ program (Dix et al. 2007; Judson et al. 2014; Kavlock et al. 2012) and the interagency Tox21 collaboration (Tice et al. 2013) involving the EPA, the U.S. Food and Drug Administration (FDA), the National Institutes of Health (NIH), and the National Toxicology Program (NTP). These two programs include assays that measure multiple steps of the ER and AR signaling pathways following the typical nuclear receptor activation process (Judson et al. 2018). Note that the ToxCast™ program also includes a steroidogenesis assay in the H295R cell line, measuring perturbations levels of multiple hormones, including testosterone (Haggard et al. 2018; Karnaus et al. 2016). However, the current work does not use this information.

In vitro HTS assays are faster and more cost-effective than traditional in vivo toxicity testing, and they avoid the ethical concerns associated with animal tests. However, no single assay is currently sufficient (due to lack of accuracy, cytotoxicity, solubility issues, etc.) to screen all classes of chemicals for a given molecular target and activity mode (e.g., agonist or antagonist) for an accurate evaluation of potential harm to human health and the environment (Judson et al. 2015, 2018). Therefore, it has been necessary to develop batteries of assays covering various and steps of the ER and AR pathway signaling processes, which increases the time and costs that are necessary to run and analyze the data. Also, such assays are not applicable to chemicals that are still in the molecular development and optimization phases. Thus, a prioritization of existing chemicals and a virtual screening of new ones being designed are necessary steps to provide knowledge about chemicals with little or no known experimental data (Judson et al. 2018). With the recent technological advances in computational resources and machine learning algorithms, in silico approaches, such as quantitative structure–activity relationships (QSARs), are particularly appealing as fast and economical alternatives for their ability to accurately predict toxicologically relevant end points (Dearden et al. 2009; Worth et al. 2005). These methods are based on the concept that similar structures are associated with similar biological activity (Hansch and Fujita 1964).

The use of computational methods to screen and prioritize chemicals for endocrine activity has been already initiated at the EPA’s National Center for Computational Toxicology (NCCT), the EPA Office of Science Coordination and Policy, and the NTP Interagency Center for Evaluation of Alternative Toxicological Methods (NICEATM), with a special focus on ER and AR. Starting with ER, a total of 18 ToxCast™ and Tox21 in vitro assays targeting the main estrogen-signaling steps (three cell-free radioligand binding assays; six dimerization assays using both ERα and ERβ; two DNA binding assays; two RNA transcription assays; two agonist-mode protein expression assays; two antagonist-mode protein expression assays; and one cell proliferation assay) were run on a library of 1,855 ToxCast™ chemicals (Richard et al. 2016). Then, a mathematical pathway model combined the results into a unique area under the curve (AUC) score [0–1] overcoming the limitations of single assays (assay interference and cytotoxicity) as an estimate of ER pathway activity (Judson et al. 2015). These in vitro model scores were then used by a consortium of 40 scientists from 17 international research groups, coordinated by NCCT, in the framework of the Collaborative Estrogen Receptor Activity Prediction Project (CERAPP) (Mansouri et al. 2016a) to develop models for ER binding, agonist, and antagonist activity. A total of 48 QSAR and dock- ing predictive models were developed, which were evaluated using an external set from the literature and subsequently combined into consensus models. The consensus models were then used to virtually screen a library of 32,464 unique chemical structures compiled from different lists of interest to the EPA, which identified approximately 4,000 chemicals with evidence of ER activity (Mansouri et al. 2016a). CERAPP demonstrated the possibility of screening large lists of environmentally relevant chemicals in a fast and accurate way by combining multiple modeling approaches to overcome the limitations of single models (Mansouri et al. 2016a). In addition to the collected data and the screened list of chemicals, this project also provided a successful collaboration example to follow for using large amounts of high-quality data in model-fitting and rigorous procedures for the development, validation, and use of efficient and accurate methods to predict human or environmental toxicity while
reducing animal testing. Its workflows are now being applied to other collaborative modeling projects for different toxicological end points such as acute oral systemic toxicity (Kleinstreuer et al. 2018b).

Here, we describe a modeling project that aimed to virtually screen chemicals for their potential AR activity. The template process established by CERAPP was adopted to tackle the AR modeling project. First, a multisassay AR pathway model was developed based on the results of 11 assays covering the androgen signaling pathway and combining the in vitro results into an AUC score representing the whole AR activity to mimic the in vivo results (Kleinstreuer et al. 2017). These assays were run on the same initial library of 1,855 ToxCast™ chemicals that the ER assays were run on, and the developed pathway model was validated using reference chemicals with known in vitro results from the literature (Kleinstreuer et al. 2017) and further compared with a set of chemicals with reproducible results in vivo (Browne et al. 2018; Kleinstreuer et al. 2018a). Note that the goal of this project is to predict in vitro AR activity. There is significant discrepancy between in vitro AR activity and the results of the in vivo Hersberger assay, especially for antagonist mode. However, as demonstrated by Kleinstreuer et al. (Kleinstreuer et al. 2018a), most of these discrepancies are due to the in vivo activity occurring at internal concentrations well above the upper limit of testing in the in vitro assays (100 μM). The resulting AR pathway activity AUC scores were used as a training set in a large AR modeling consortium called the Collaborative Modeling Project for Androgen Receptor (CoMPARA). Collaborators from 25 international research groups (Supplemental Material S1) contributed a total of 91 qualitative and quantitative predictive QSAR models for binding, agonist, and antagonist AR activities. The total list of chemicals that CoMPARA participants screened using their models comprised 55,450 chemical structures, including CERAPP chemicals and ToxCast™-generated metabolites (Leonard et al. 2018; Pinto et al. 2016). These predictions were evaluated using curated literature data sets and then combined into binding, agonist, and antagonist consensus models. Both CERAPP and CoMPARA projects were collaborations between the participants, aiming to build the best collective consensus rather than competing for the best single model. We also describe the procedure of extending CERAPP and CoMPARA consensus models beyond their original lists that was used in the screening of the entire EPA DSSTox database (https://comptox.epa.gov/dashboard; Gruilke et al. 2019) and other chemicals of interest that are structurally similar to the initial lists (Williams et al. 2017).

Figure 1. Workflow of the project defining the major steps and the different data sets used for training, evaluation, and prediction.
Materials and Methods
CoMPARA followed the template defined by the CERAPP research effort, taking into account the learnings and best practices to update scripts and workflows applied to AR data (Mansouri et al. 2016a). The steps were as follows:

1. Preparation of the AR pathway data as derived from the biological model using the 11 ToxCast™ assays (Kleinstreuer et al. 2017), which formed the basis of a common training set.
2. Compilation of the prediction set, which was the list of chemicals to be screened by the collaborators.
3. Collection and curation of an external evaluation set, which comprised data extracted from the literature to be used for evaluating the predictive ability of the models (mostly for verification purposes and not to compare models), performed in parallel with the model building efforts.
4. Model evaluation process and generation of consensus predictions once all models were submitted.
5. Validation and extension of the consensus models for future screening procedures.

Figure 1 represents the workflow of the project and the genesis of the different chemical sets (training, evaluation, and prediction).

Data Sets
A number of different data sets were created and applied at various stages of the project, described in detail below. First, a common training set was compiled and provided to the participants to fit their models. Subsequently, participants were provided a prediction set consisting of the list of chemicals to be screened using their models. While modelers were fitting the training set, other data were being collected and curated from the literature to be used as an evaluation set to assess the predictive ability of the models.

Training Set, the ToxCast™ AR Pathway Model
As was done for ER, the AR in silico efforts started with the development of a multiassay in vitro model covering the signaling pathway. A battery of 11 ToxCast™/Tox21 in vitro assays were selected: three receptor binding, two cofactor recruitment, one RNA transcription, three agonist-mode protein production, and two antagonist-mode protein production (Judson et al. 2018; Kleinstreuer et al. 2017). One of the antagonist mode assays was run with two different concentrations of the stimulating ligand to provide confirmation data for receptor-mediated activity and to further distinguish true antagonist pathway activity from cell stress or cytotoxicity-mediated loss of function. The 1,855 ToxCast™ chemicals were run through these assays and the resulting data were combined using a mathematical model to yield a unique AUC score for AR agonist and antagonist activity (Kleinstreuer et al. 2017). This score was used in combination with a confidence score derived from confirmation assay data (using a higher concentration of the activating ligand to characterize competitive binding) and a bootstrapping procedure so that chemicals with an AUC of at least 0.1 (corresponding to activity concentrations up to 100 μM) were considered actives, chemicals with AUC less than 0.001 were considered inactives, and the remaining chemicals were considered inconclusive (Kleinstreuer et al. 2017). This model, accounting for assay interference and cytotoxicity, was validated using 54 reference chemicals from the literature (Kleinstreuer et al. 2017). Because the AUC scores are available only for agonist and antagonist activity, for the purpose of this project (as well as in CERAPP previously) a chemical was considered to be a binder if it were either an active agonist or antagonist.

This high-quality data set covering the AR signaling pathway, however, contains a very low fraction of actives: approximately 10% antagonists and only 2% agonists. Because this bias toward inactives can be challenging for modelers, a literature search was conducted to identify additional actives. However, with the lack of sources matching our data (whole AR pathway activity), a list of only 15 active chemicals (13 agonists and 2 antagonists) collected from DrugBank was added to the data set (Wishart et al. 2008). Being pharmaceuticals, these chemicals were assigned an AUC score of 1 as strong actives, even though they were not tested in the 11 ToxCast™ assays.

The KNIME (Konstanz Information Miner) chemical structure standardization workflow developed for CERAPP was applied to the available structures and generated a total of 1,688 unique, organic, desalted QSAR-ready structures (Mansouri et al. 2016a; McEachran et al. 2018).

The resulting three data sets (binding, agonist, and antagonist; Table 1) were provided to the modelers in three separate two-dimensional structure data files (2D SDF V2000) with QSAR-ready standardized structures in both Simplified Molecular Input Line Entry System (SMILES) and molecular format with atom coordinates (MOL) formats. The three-dimensional (3D) structures were also generated by energy minimization using MMFF94 force field and provided as 3D structure data file (sdf) (V2000) files. In both the 2D and 3D files, area under the curve (AUC) values were provided with the corresponding activity class (binary) and converted concentration values (AC50) indicating potency (Judson et al. 2015). In addition to the structures and data, each chemical was also given an associated CASRN, DTXSID identifier, preferred name, standardized InChI code, and a hashed InChI key of the Quantitative Structure-Activity Relationship (QSAR)-ready structures.

Each one of these data sets (available in Supplemental Material S2) could be used to build either continuous models predicting AUC values or categorical models predicting active and inactive classes. A list of chemicals in these three data sets could be considered false negatives thus, could be removed by the participants during the modeling procedures. These potentially inconclusive chemicals (available in Supplemental Material S3) consisted of 21 chemicals from binding, 8 from agonist, and 14 from antagonist. Further filtering based on clustering or other structure-based analysis was recommended to reduce the number of inactives and to decrease the bias between the two classes. The ToxCast™-derived data sets (provided in SDF file format), as well as the removed chemicals were made available for download at https://doi.org/10.23645/epacomptox.10321697.

Prediction Set, Structure Collection, and Curation of Lists of Interest
The initial CERAPP list was compiled from a library of over 50,000 chemicals that humans are potentially exposed to from lists of toxicological and environmental chemicals of interest. The main lists included the EPA’s Chemical and Product Categories database (U.S. EPA 2014a), the first version of the EPA’s Distributed Structure-Searchable Toxicity Database (DSTox) (Richard et al. 2016; Richard and Williams 2002), and the Canadian domestic substance list (Environment Canada 2002). This library contained a total of 42,679 chemicals with known organic structures that, after QSAR-ready standardization procedure and duplicates removal, was collapsed to 32,464 unique structures known as the CERAPP list (Mansouri et al. 2016a).

Table 1. Training set chemicals for binding, agonist and antagonist data sets.

<table>
<thead>
<tr>
<th>Number of</th>
<th>Binding</th>
<th>Agonist</th>
<th>Antagonist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>1,662</td>
<td>1,659</td>
<td>1,525</td>
</tr>
<tr>
<td>Actives</td>
<td>198</td>
<td>43</td>
<td>159</td>
</tr>
<tr>
<td>Inactives</td>
<td>1,464</td>
<td>1,616</td>
<td>1,366</td>
</tr>
</tbody>
</table>
In CoMPARA, in addition to the lists included in CERAPP, we used the European inventory of existing commercial chemical substances (EINECS) containing ~ 60,000 chemicals as a list of interest for in silico screening. We also incorporated ToxCast™ metabolites in the prediction set that had been generated as part of related ER studies (Leonard et al. 2018; Pinto et al. 2016). The goal of including metabolites in the CoMPARA project was to understand the effect of xenobiotic metabolism, which is lacking in most in vitro assays. For ER screening efforts, this step was conducted post CERAPP in two different studies generating a total of 15,406 metabolite structures for ToxCast™ parent chemicals using ChemAxon Metabolizer (discontinued 2018) (ChemAxon, Ltd.) (Leonard et al. 2018; Pinto et al. 2016). After QSAR-ready standardization and removal of duplicates, the CoMPARA list consisted of 55,450 QSAR-ready structures with unique CoMPARA integer IDs, including 6,592 nonredundant metabolite structures. This list matches 63,848 original (pre-QSAR-ready) chemical structures in the EPA’s DSSTox database, excluding the metabolites.

The CoMPARA chemical prediction set was provided in 2D and 3D SDF files. Data provided for each chemical included CoMPARA IDs; structures in SMILES, MOL, and InChI code; and hashed InChI key formats for all chemicals, CASRN names, and DSSTox DTXSIDS were also provided when available. This list of chemicals (identifiers and structures in SMILES format) is provided in Supplemental Material S4. The two QSAR-ready files as well as the original (prestandardization) structures file were made available for download at https://doi.org/10.23645/epacomptox.10321697.

Evaluation Set, Literature Search, and Curation
To assess the developed models and their predictivity, an evaluation set with new chemicals (nonoverlapping with the training set) is required. Ideally, this new set would be the result of the same mathematical pathway model combining the 11 assays as that used for the training set. Because such a data set was not available, we decided to use data from the literature. Because the project was not a competition and the goal of this step was not to provide an in-depth comparison of the models, literature data could still be used to provide quality assessment and check for errors.

Large amounts of experimental data are available in the PubChem repository and related data sources. However, such public sources of chemical-biological data have varying levels of quality control, so additional curation and standardization are necessary (Williams and Ekins 2011). The EPA’s NCCT collected and curated PubChem data (64 sources), restructured it, and mapped the bioactivity values to related biological targets. In this effort, we started with ~ 80,000 experimental values for AR activity, which mapped to about ~ 11,000 chemicals that we grouped by modality (agonist, antagonist) and hit call (active, inactive). To improve the consistency between the different PubChem entries and to add binding modality, three rules were applied:

• In the case of multiple records for a test chemical, a minimum concordance of two out of three assay results was required to assign a positive activity score.

• An active agonist or antagonist was considered a binder.

• Inactive agonists and antagonists were considered nonbinders.

The KNIME standardization workflow referenced earlier was applied to the chemical structures (Mansouri et al. 2016a; McEachran et al. 2018). After removing ToxCast™ chemicals (used for the training set), the generated standard InChI codes matched 7,281 chemicals from the CoMPARA list (prediction set). This list of 7,281 chemicals, with associated data extracted from the literature, was used as the evaluation set. The removed ToxCast™ chemicals were mostly associated with ToxCast™ data only. The evaluation set chemicals were split into three data sets based on the available experimental data. The resulting lists included 4,839 structures for agonist, 4,040 for antagonist, and 3,882 for binding. The numbers of active and inactive structures are summarized in Table 2.

AC50 values (μM) were available for 405 chemicals in the binding data set, 167 chemicals in the agonist data set, and 340 chemicals in the antagonist data set. This process of preparing the evaluation set was conducted in KNIME. These three data sets (identifiers and structures in SMILES format) are provided in Supplemental Material S5. The SDF files have been made available for download at https://doi.org/10.23645/epacomptox.10321925.

Participants and Modeling Methods
CoMPARA participants included a total of 70 scientists from 25 international research groups representing academia, governmental institutions, and industry (See Supplemental Material S1), including 15 groups that were involved in the related CERAPP project. The participating groups were located in 11 different countries. The modelers were encouraged to use the provided training set and, preferably, apply free and open-source software tools that included detailed descriptions of the used methods and the employed applicability domain assessment. However, the application of proprietary commercial tools was also allowed. The different molecular descriptor calculation tools and the various modeling approaches employed are summarized in Tables 3 and 4. Some of the groups collaborated with each other to submit common models. To produce more balanced data, most participants applied under-sampling approaches to reduce the number of inactive chemicals. For further details on the methods and approaches, see the full list of files submitted by the individual groups at https://doi.org/10.23645/epacomptox.10321982. Certain participants developed additional approaches applied by the participating groups.

<table>
<thead>
<tr>
<th>Number of Binding</th>
<th>Agonist</th>
<th>Antagonist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actives</td>
<td>453</td>
<td>167</td>
</tr>
<tr>
<td>Inactives</td>
<td>3,429</td>
<td>4,672</td>
</tr>
<tr>
<td>Total</td>
<td>3,882</td>
<td>4,839</td>
</tr>
</tbody>
</table>

Note: —, No data.

Approaches are sorted alphabetically by acronym.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Approach</th>
<th>Reference to specific method, if published</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>Artificial neural networks</td>
<td>—</td>
</tr>
<tr>
<td>ASNN</td>
<td>Associative artificial neural networks</td>
<td>Tetko 2002; Tetko and Tanchuck 2002</td>
</tr>
<tr>
<td>DF</td>
<td>Decision forest</td>
<td>Hong et al. 2005, 2004; Tong et al. 2003; Xie et al. 2005</td>
</tr>
<tr>
<td>DT</td>
<td>Decision trees</td>
<td>Berk 2008; Mazanetz et al. 2012</td>
</tr>
<tr>
<td>GBM</td>
<td>Gradient boosting method</td>
<td>Martin et al. 2008</td>
</tr>
<tr>
<td>HC</td>
<td>Hierarchical clustering</td>
<td>Cover and Hart 1967; Kowalski and Bender 1972</td>
</tr>
<tr>
<td>kNN</td>
<td>k nearest neighbors</td>
<td></td>
</tr>
<tr>
<td>LDA</td>
<td>Linear discriminant analysis</td>
<td></td>
</tr>
<tr>
<td>MLR</td>
<td>Multilinear regression</td>
<td></td>
</tr>
<tr>
<td>NB</td>
<td>Naïve Bayes</td>
<td>Murphy 2006</td>
</tr>
<tr>
<td>PLS</td>
<td>Partial least squares</td>
<td>Wold et al. 2001</td>
</tr>
<tr>
<td>PLSDA</td>
<td>Partial least squares discriminative approach</td>
<td>Frank and Friedman 1993; Nouwen et al. 1997</td>
</tr>
<tr>
<td>RBF</td>
<td>Radial basis function</td>
<td>Zakharov et al. 2014</td>
</tr>
<tr>
<td>RF</td>
<td>Random forest</td>
<td>Brennan 2001</td>
</tr>
<tr>
<td>SCR</td>
<td>Self-consistent regression</td>
<td>Lagunin et al. 2011</td>
</tr>
<tr>
<td>SVM</td>
<td>Support-vector machines</td>
<td>Cortes and Vapnik 1995</td>
</tr>
</tbody>
</table>

Table 2. Evaluation set chemicals for binding, agonist, and antagonist data sets.

Table 3. Modeling approaches applied by the participating groups.
models and published them or expressed their intent to publish them as separate manuscripts.

**Evaluation Procedure**

To ensure consistency and repeatability of the results, all molecular structures associated with the chemicals in the three data sets described above (training set, prediction set, and evaluation set) were processed using the same standardization workflow. This workflow was designed so that chemicals with the same standardized QSAR-ready structures would automatically have the same computational predictions (Fourches et al. 2010, 2016; Mansouri et al. 2016a). Thus, chemicals from the different data sets could be matched using their respective QSAR-ready standard InChI codes.

After all predictions were made on the prediction set, the overlap was designed so that chemicals with the same standardized structures associated with the chemicals in the three data sets could be matched using their respective QSAR-ready standard InChI codes. The main evaluation criteria were:

- **Goodness-of-fit**: statistics of the training set.
- **Predictivity**: statistics of on the evaluation set.
- **Robustness**: balance between goodness-of-fit and predictive ability.

Each of these three criteria was assigned a weight resulting in a score ($S$) ranging from 0 to 1:

$$S = 0.3 \times \text{(Goodness of fit)} + 0.45 \times \text{(Predictivity)} + 0.25 \times \text{(Robustness)}.$$  \hspace{1cm} (1)

This score was not intended to rank the models but was designed by the organizers of the consortium mainly to evaluate the models’ predictivity and provide a rational basis to combine the predictions into a consensus in a later step. The weights were assigned to the different components in a way to give priority to the predictive ability on the evaluation set but not too high in comparison with the training set statistics because of the difference between the two sets. The robustness or balance between the two was given the third rank but just slightly less weight than the goodness of fit because it highlights overfitting, which is almost as important a factor as fitting. However, a slightly different weight would most probably lead to the same results. To ensure equal contribution from the participants, the evaluation score accounted neither for the fraction of predicted chemicals nor the coverage of the AD, if provided.
For the qualitative models, this general formula was translated into commonly used classification parameters as discussed below. However, for the quantitative models, the predictions based on the training set data (ToxCast™ AUC values) were different from those of the evaluation set data (AC50 values). To ensure consistency and validity of the evaluation procedure, the qualitative and quantitative models and predictions were processed separately as explained below.

**Qualitative Evaluation Procedure**

The categorical predictions were evaluated using statistical indices commonly proposed in the literature (Ballabio et al. 2018). These indices are calculated from the confusion matrix, which collects the number of samples of the observed and predicted classes in the rows and columns, respectively. The classification parameters are defined using the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The resulting parameters were the balanced accuracy (BA), specificity (Sp), and sensitivity (Sn) calculated as follows:

The BA is given by:

$$BA = \frac{(Sn + Sp)}{2},$$  \hspace{1cm} (2)

where the sensitivity (Sn), or true positive rate (TPR) is given by:

$$Sn = \frac{TP}{TP + FN},$$ \hspace{1cm} (3)

and the specificity (Sp), or true negative rate (TNR) is given by:

$$Sp = \frac{TN}{TN + FP}.$$ \hspace{1cm} (4)

For classification models, not only is the average of the Sn and Sp explained by the BA important, but also the balance between them. Therefore, to fully assess the predictivity of the models, the three criteria are included in the general scoring function $S$, calculated as follows:

Goodness of fit = $0.7 \times (BA_{Tr}) + 0.3 \times (1 - |Sn_{Tr} - Sp_{Tr}|)$, \hspace{1cm} (5)

Predictivity = $0.7 \times (BA_{Eval}) + 0.3 \times (1 - |Sn_{Eval} - Sp_{Eval}|)$, \hspace{1cm} (6)

Robustness = $1 - |BA_{Tr} - BA_{Eval}|$, \hspace{1cm} (7)

where $Tr$ stands for training set and $Eval$ stands for evaluation set, attributing weight not only to the BA but also to the balance between Sn and Sp to account for the reliability of the model in predicting actives as well as inactives.

**Quantitative Evaluation Procedure**

Active chemicals with available quantitative information (AC50 values) from the mined literature sources (405 chemicals in the binding data set, 167 chemicals in the agonist data set and 340 chemicals for the antagonist data set) were considered for this step to evaluate the quantitative models’ predictivity. The analysis of the quantitative results conducted during the CERAPP project showed some differences between the AC50 values collected from the literature and the AC50 values converted from the predicted AUC scores. These differences include the fact that the AUC scores represented the results of multiple assays that were combined to overcome assay interference and cytotoxicity, whereas the literature data is a one source per assay most of the time. In addition, the ToxCast™ assays’ limiting dose of 100 μM makes these assays insensitive to “very weak” actives that are reported in the literature to have AC50 values beyond that threshold. Thus, to conduct a quantitative evaluation of the predictions using the literature data without underestimating the accuracy of the predictions, the two types of results needed to be converted to a more consistent data type. The multiclass approach presented in this work converting both the literature data and the predicted AUC values to five categories with approximately similar potencies was built on the CERAPP approach (Mansouri et al. 2016a). This approach is commonly used in the QSAR field to predict end points that are hard to model on a continuous scale and to avoid underestimating predictivity (Benigni 2003; Dunn 1990; Kowalski 2013; Waterbeemd 2008). This approach was applied in CoMFA to avoid the same problem when evaluating the models that were trained on the ToxCast™-based AR pathway model (AUC values) using literature data. Both literature (evaluation set chemicals with quantitative information) and predicted data sets were categorized into five potency activity classes: inactive, very weak, weak, moderate, and strong [example reference chemicals with different potency levels are available in Judson et al. and Kleinstreuer et al. (Judson et al. 2015; Kleinstreuer et al. 2017)]. These five classes were used to evaluate the quantitative predictions.

The thresholds determined in the CERAPP project were applied to the concentration-response values (AC50) from the literature:

- Strong: Activity concentration below 0.09 μM
- Moderate: Activity concentration between 0.09 and 0.18 μM
- Weak: Activity concentration between 0.18 and 20 μM
- Very Weak: Activity concentration between 20 and 800 μM
- Inactive: Activity concentration higher than 800 μM

For the training set, the AUC values were converted into five potency classes based on the following thresholds:

- Strong: AUC equal to or above 0.75
- Moderate: AUC between 0.75 and 0.65
- Weak: AUC between 0.65 and 0.25
- Very weak: AUC between 0.25 and 0.09
- Inactive: AUC below 0.09

Although the ToxCast™ single assays were limited to a maximum concentration of 100 μM, similar to CERAPP, active chemicals with an AUC score below 0.25 are considered “very weak.” However, for the lack of chemicals in the 0.25 to 0.5 AUC range (weak in CERAPP), this arbitrary range for weak actives was extended to 0.65. The five classes were assigned labels from 1 (inactive) to 5 (strong), and then the models were evaluated as multiclass categorical models in binding, agonist, and antagonist modes separately. The above-mentioned formulas for calculating Sn, Sp, and BA were applied to each of the classes, and then the average values (for the five classes) were inserted into the scoring function.

**Consensus Modeling**

After being evaluated separately according to the defined strategy, each model was given a score ($S$) for predictions within its AD. This score was used in the consensus modeling step as a weighting factor. Using these weights, the predictions within the AD of the submitted models were combined into a single consensus model separately for each modality (binders, agonists, and antagonists). For each chemical in the prediction set, the consensus category was decided by the weighted majority rule: the chemical was assigned the class with the highest average score of the models predicting it (class with the highest average score was selected). This average score excluded the models that did not provide a prediction for the chemical in question.

The consensus model predictions were evaluated using the same evaluation set and procedure used to evaluate the individual
models, and their performances were compared to the single models. Analyses of the accuracy trends and concordance (fraction of consistent predictions) among the predictions of the different models were also conducted. Considering only the models that provided predictions, the sum of the concordance among models for actives and inactives should be equal to 1.

**Generalization of the Consensus and Implementation in OPERA**

To use the consensus models beyond the initial prediction set, the combined predictions were used to train generalized models capable of replicating the original consensus. This procedure was achieved by applying a weighted k-nearest-neighbor (kNN) approach to fit the classification models based on the majority vote of the nearest neighbors. This approach has the advantage of resembling read-across, which is a broadly accepted data-gap filling tool in regulatory agencies (Ball et al. 2016; Patlewicz et al. 2017). In addition, kNN models can also satisfy the five OECD principles for QSAR modeling due to their nonambiguous algorithm, high accuracy, and interpretability (Buttrey 1998; Cover and Hart 1967). Furthermore, being distance-based (dissimilarity), the weighted kNN approach fits the purposes of extending the consensus predictions to new chemicals and providing the exact same prediction for the chemicals that already have a consensus model prediction. This goal is achieved by considering the first nearest-neighbor prediction if the distance is zero (100% similarity). An applicability domain index is also provided to assess the similarity of the predicted chemical to the nearest neighbors.

PaDEL (version 2.21) and CDK (version 2.0) were used to calculate two-dimensional molecular descriptors (Guha 2005; Yap 2011). Because PaDEL uses a previous version of CDK (1.5), overlapping descriptors between it and CDK2 were excluded. The union of the PaDEL descriptors (1,444) and CDK2 (287) resulted in a total of 1,616 variables that were later filtered for low variance and missing values.

Here, kNN was coupled with genetic algorithms (GA) to select a minimized optimal subset of molecular descriptors. GA begins with an initial random population of chromosomes, which are binary vectors representing the presence or absence of molecular descriptors. An evolutionary process is then simulated to optimize a defined fitness function and new chromosomes are obtained by coupling the chromosomes of the initial population with genetic operations such as crossover and mutation (Ballabio et al. 2011; Leardi and Lupiáñez González 1998).

The best models were selected and implemented in OPERA, a free and open-source suite of QSAR models (Mansouri et al. 2016b, 2018). Both OPERA’s global and local AD approaches, as well as the accuracy estimation procedure, were applied to the predictions (Mansouri et al. 2018). The global AD is a Boolean index based on the leverage approach for the whole training set, whereas the local AD is a continuous index in the [0–1] range based on the most similar chemical structures from the training set (Mansouri et al. 2018).

**Results and Discussion**

**Received Models**

There was a total of 91 models submitted by the participating groups. Each submission consisted of predictions for the full or fraction of the prediction set using one or multiple models, as well as the related documentation with various levels of detail. All submitted results for the prediction set are provided in Supplemental Material S6. The full list of files submitted by the participants is available at https://doi.org/10.23645/epacomptox.10321982. The submissions included categorical and continuous predictions from binding, agonist, and antagonist models as shown in Table 5.

The number of categorical models submitted greatly exceeded the number of continuous models. This difference is due to the difficulty of modeling the low number of AUC values greater than zero in the provided training set (ToxCast™ data). All participating groups provided at least one binding model. Thus, the number of binding models is higher than the number of either agonist or antagonist models. This higher number is also due to the biased training data, which included low numbers of active agonists and antagonists. As described above, the number of active binders is the union of both active agonists and antagonists.

**Results of the Evaluation Procedure**

The evaluation procedure described above was applied to the categorical and continuous predictions provided by the participants. The goal of this step was to assess the accuracy of the predictions that, in a later step, were combined into the consensus models. Thus, the evaluation procedure was not designed to reflect the uneven coverage of the prediction set, and application of an AD was encouraged.

The first application of this evaluation procedure revealed models that suffered from mishandling of data that might occur during the modeling process. Such errors led to a severe decrease in prediction accuracy and included mismatches between the IDs of the prediction set chemicals and their associated predictions, as well as misinterpretation (inverting or mismatching) of the different fields contained in the provided files (training and prediction sets) introduced by certain automated procedures. Because the goal of the project was not to compete, but rather to collaborate, the submitters of the models with such issues were notified to correct them to allow for a better contribution toward the consensus. The final, corrected submissions were used to produce the statistics provided in Supplemental Material S7. The results of the evaluation procedure, discussed below, are also available at https://doi.org/10.23645/epacomptox.10321994.

**Qualitative Models**

As mentioned earlier, all participating groups built categorical binding models. Furthermore, certain groups, such as IRFMN and ATSDR, collaborated with others to provide additional models (Manganelli et al. 2019). For binding activity, ~85% of the models achieved a BA above 0.8 for the training set, and about 70% of them achieved a BA above 0.7 for the evaluation set. On average, the BA of most models decreased ~15% for the evaluation set relative to the training set. We consider this decrease to be negligible based on the differences between the training set data (which are based on the AUC values of the ToxCast™ AR pathway model that combines multiple assays) and the evaluation set data (which are consensus hit calls from the literature or rely on only one record for a particular chemical with unknown reproducibility). With this high predictivity and balance between training and evaluation set performance increasing their robustness, ~75% of the binding models reached a score of 0.75.

The agonist models showed a higher performance in comparison with the binding models. Indeed, all agonist models had a training BA above 0.8, and 95% of them achieved a BA above

<table>
<thead>
<tr>
<th>Table 5. Summary of the submitted models.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Categorical</td>
</tr>
<tr>
<td>------------------------</td>
</tr>
<tr>
<td>Binding</td>
</tr>
<tr>
<td>Agonist</td>
</tr>
<tr>
<td>Antagonist</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>
0.75 on the evaluation set. This performance brings the average difference between training and test BA down to 0.1, indicating lower risk of overfitting. Most (∼95%) of the agonist models achieved a score above 0.85 (Figure 2).

Although the data sets contained more active antagonists in comparison with agonists, the general performance of the antagonist models was inferior. This inferiority was reflected in the evaluation set performance, because only two models reached a BA of 0.75, which affected their robustness (0.24 average difference between the training and evaluation BAs). However, the average and median scores reached 0.78 and 0.79 respectively, which shows high general performance of the models (Figure 2).

Most of the submitted categorical models (agonist, antagonist, and binding) provided predictions for the majority of the prediction set chemicals (>90%). Some of the participants who submitted more than one model, such as NCATS and UNIMIB, opted for a low coverage with high accuracy and a high coverage with lower accuracy. Additional details about the performance of the models is provided in Supplemental Material S7.

### Quantitative Models

The quantitative models were converted into multiclass categorical models as described in the Methods section. The overall performance of the thirteen models across the three modalities (agonist, antagonist, and binding) was lower than that of the binary categorical models. The binding models performed a bit better than the agonist and antagonist models. Indeed, four binding models out of five obtained a score above 0.6, whereas only one agonist and one antagonist model performed that well (Figure 3). The predictive accuracy of these models was also assessed on the five classes separately (details available in Supplemental Material S7). This analysis showed that most models exhibited BAs of approximately 0.5 for the five classes, with the binding models exhibiting slightly better performance (0.7–0.78) in identifying inactives.

### Consensus Modeling

Based on their low number and average performance in comparison with the categorical models, the recommendation would be that the continuous models should be used individually. A continuous consensus model can be derived only from a more concordant set of models. Thus, for the sake of accuracy and consistency of the predictions, only the categorical models were considered for the consensus step. Before combining the categorical predictions into a consensus, we checked the coverage and concordance among the models. As shown in Figure 4, all chemicals in the prediction set are covered by at least 11 models. Moreover, most chemical structures can be predicted by 18–20 agonist and antagonist models. For binding, most chemicals were predicted by 28–31 models. This high coverage provides a good basis for the consensus model and strengthens the statistical relevance of the combined predictions.

The concordance among the models is also an important criterion for combining the predictions. In fact, chemicals predicted with high concordance among numerous models built using different modeling approaches can inform on accuracy (Mansouri et al. 2016a). Figure 5 shows that most binding, agonist, and antagonist categorical predictions are at least 90% concordant. Because most models were associated with comparable scores, the average score used to categorize chemicals was largely in

![Figure 2](image)

**Figure 2.** Scores of the categorical binding (black), agonist (white) and antagonist (gray) models based on the evaluation set and the scoring Equation 1.

![Figure 3](image)

**Figure 3.** Scores of the continuous binding (black), agonist (white) and antagonist models based on the evaluation set and the scoring Equation 1 (See Supplemental Material 1 for groups’ abbreviations).
agreement with model concordance; i.e., the average score for actives was high when the concordance among the models with active predictions was high, and vice versa. A few exceptions were noted when model concordance was around 0.5, which indicated that only one or two models were driving the classification. Thus, based on these statistical observations about the concordance between the models, it can be concluded that it is possible to combine the categorical predictions into consensus agonist, antagonist, and binding predictions.

**Consensus Predictions**

The predictions from the binding, agonist, and antagonist categorical models were at first combined independently based on the calculated scores. Because the participants provided uneven fractions of the prediction set, the resulting predictions for each of the 55,450 chemical structures were based on different contributing models. Thus, predictions from the same model can be associated with different weights across the prediction set. After generating the consensus predictions for the whole prediction set, the same evaluation procedure described previously was applied to each of the single models. The resulting statistical details are reported in Supplemental Material S7 as CONSENSUS_1. All obtained parameters, including the BA for the training and evaluation sets as well as the corresponding score, were around the top values obtained for the single models.

In a second step, to improve consistency among the agonist, antagonist, and binding predictions across the whole set of 55,450 chemicals, the following set of rules were applied:

- If a chemical i was predicted to be an active agonist or antagonist with a concordance below 70%, but an inactive binder with a concordance above 70%, it was considered an inactive agonist or antagonist, respectively.
- If a chemical i was predicted to be an active agonist or antagonist with a concordance above 70%, but an inactive binder with a concordance below 70%, it was considered an active binder.

After applying these corrections to the consensus predictions, the total numbers of actives and inactives changed slightly for the three modalities (Table 6). However, the overall statistics as reported as CONSENSUS_2 in Supplemental Material S7 remained almost unchanged. This result is because most of the corrected predictions are not included in the evaluation set.

The evaluation results of the final consensus predictions are reported in Table 7. The final consensus for the whole prediction set (identifiers and structures in SMILES format) are provided in Supplemental Material S8, and the SDF files are available at [https://doi.org/10.23645/epacomptox.10322012](https://doi.org/10.23645/epacomptox.10322012).

Because the training and evaluation sets were designed such that active agonists and antagonists were considered active binders, the corrected predictions can be considered more consistent with these two data sets. However, chemicals with inactive binding prediction and active agonist or antagonist that are all below 70% concordance were not changed. Also, certain chemicals were predicted to be active binders but inactive in both agonist and antagonist modalities. This circumstance was also noticed in the CERAPP predictions and was resolved by classifying these substances as low potency binders (Mansouri et al. 2016a). Similarly, certain chemicals were predicted to be active agonists and antagonists simultaneously. Such chemicals were also present in the ER and AR ToxCast™ data, as well as CERAPP predictions, and were considered to be strong in one modality but weak in the other.

Table 7 shows a noticeable drop in Sn and in the associated BA result for the whole evaluation set with its five potency classes. However, this drop does not indicate low performance of the single models or the resulting consensus predictions. It is more of an indication of differences between the two data sets. Usually, assay technology and other experimental differences can cause such discordance for certain chemicals. However, in this case, the low sensitivity of the consensus model on the evaluation set that led to the drop in accuracy in comparison with the training set is most probably due to the differences in the ranges of testing between ToxCast™ and the literature sources. Another cause of the difference between the two data sets is that the ToxCast™ data is a result of multiple assays that were combined based on a pathway model designed to avoid assay interference and cytotoxicity leading to false positives. This result can occur when chemicals are tested at

![Figure 4](image-url) Histogram showing the distribution of the number of binding (black), agonist (white) and antagonist (gray) models covering the prediction set (minimum of 11 models for agonist and antagonist and 20 for binding).

![Figure 5](image-url) Histogram showing the distribution of the concordance of the binding (black), agonist (white) and antagonist (gray) single models.

| Table 6. Total numbers of predicted actives and inactives before and after corrections. |
|---------------------------------|---------------|---------------|---------------|---------------|---------------|
|                                 | Binding Actives | Binding Inactives | Agonist Actives | Agonist Inactives | Antagonist Actives | Antagonist Inactives |
| Before                          | 9,878          | 45,572         | 2,239          | 53,211         | 12,705         | 42,745          |
| After                           | 10,521         | 44,929         | 2,167          | 53,283         | 12,136         | 43,314          |

Environmental Health Perspectives 027002-10 128(2) February 2020
high concentrations, leading to cell stress known as a “burst of activity” from turning multiple assays on at the same time (Judson et al. 2016). Thus, it is highly probable that at least some of the very weak actives in the evaluation set reported in the literature with high AC50 are, in reality, false positives that should be discounted. Additionally, as noticed in CERAPP, the increase in the number of sources in the literature data can provide information about the repeatability of the results and thus about the accuracy. These two hypotheses were evaluated by assessing the accuracy of the models (BA, Sn, and Sp) for the evaluation set after removing the chemicals in the “very weak” potency class and the chemicals with only one source, separately. The results of this analysis are summarized in Table 8.

Table 7. Statistics of the corrected consensus predictions using the whole available evaluation set.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sn</td>
<td>0.98</td>
<td>0.65</td>
<td>0.95</td>
<td>0.74</td>
<td>1.00</td>
<td>0.61</td>
</tr>
<tr>
<td>Sp</td>
<td>0.96</td>
<td>0.90</td>
<td>0.99</td>
<td>0.97</td>
<td>0.96</td>
<td>0.87</td>
</tr>
<tr>
<td>BA</td>
<td>0.97</td>
<td>0.78</td>
<td>0.97</td>
<td>0.86</td>
<td>0.98</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Table 8. Statistics of the consensus predictions after removing the “very weak” actives from the evaluation set.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Binding (3,535 chemicals)</th>
<th>Agonist (4,406 chemicals)</th>
<th>Antagonist (3,664 chemicals)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of chemicals</td>
<td>3,407</td>
<td>4,244</td>
<td>3,353</td>
</tr>
<tr>
<td>Sn</td>
<td>0.72</td>
<td>0.81</td>
<td>0.69</td>
</tr>
<tr>
<td>Sp</td>
<td>0.90</td>
<td>0.97</td>
<td>0.87</td>
</tr>
<tr>
<td>BA</td>
<td>0.81</td>
<td>0.89</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Accuracy and Limitations: Analysis of High and Low Concordance Chemicals

The analysis of the low concordance chemicals did not reveal any particular structural similarity. This finding is understandable because the ToxCast™ chemicals, used as a training set, were purposely selected to cover a wide range of chemical classes and uses (Richard et al. 2016). Thus, it is highly improbable that a large number of models based on different machine learning approaches and molecular descriptors would have similar coverage and accuracy trends relative to specific chemical features. However, the analysis of concordance in terms of the accuracy of prediction in the evaluation set revealed that the models were more discordant for inaccurate predictions. Figure 8 shows that the concordance is generally above 90% for accurately predicted chemicals. There are certainly a number of exceptions that contradict this observation and that cannot be linked only to situations where the majority of models are generating wrong predictions but also to known differences between the training and evaluation sources. For example, hydroxyflutamide (CAS52806-53-8 and DTXSID8033562), a known antiandrogen drug, and confirmed as a strong antagonist with an AUC (ToxCast™ combined assays) score of 0.999 and a literature AC50 of 0.262 µM (U.S. EPA 2019e, 2019f; Wikipedia 2018). Hydroxyflutamide is, as expected, predicted by CoMPARA consensus to be an AR antagonist with a 0.95 concordance (20/21 antagonist models). However, in the collected literature data, hydroxyflutamide is also a very weak agonist with reported AC50 of 23.85 µM. However, in the CoMPARA consensus agonist predictions, it is considered as inactive with a concordance of 0.95 (20/21 agonist models), which is consistent with to ToxCast™ AUC score of 0.001 (U.S. EPA 2019). This example shows that the sensitivity of the CoMPARA consensus models is more similar to the AUC score assessment, which is based ToxCast™ combined assays, rather than that in reported literature, which is usually based on a
single assay and corresponding independent reference. Another similar example is bicalutamide (CAS90357-06-5 | DTXSID2022678) (U.S. EPA 2019c, 2019d; Wikipedia 2019b). However, this comparison does not mean that all very weak actives are mispredicted. Aldosterone, for example, and many others are reported in the literature as very weak antagonists with AC50 values above 60 μM and predicted by CoMPARA’s antagonist consensus as actives with concordances above 0.75 (U.S. EPA 2019a, 2019b; Wikipedia 2019a).

As previously noted, the concordance around the inactive predictions is generally high for most cases. Hence, to reveal any trends in the active predictions, a box plot for the concordance against the potency of binders was plotted for the evaluation set chemicals (Figure 9). This analysis showed that the concordance for moderate and strong binders is clearly higher than for very weak and weak ones. Similar trends were noticed for the agonist and antagonist predictions. The decreased accuracy for the low potency chemicals can explain the low sensitivity of the consensus predictions as discussed above, and the low concordance can be an indication of low accuracy. However, as Figure 8 shows, there are accurate predictions associated with low concordance and inaccurate predictions associated with high concordance. This finding is comparable to the AD assessment, which helps provide the user with context based on the assumption that predictions in the AD are generally more accurate than those outside the AD. The AD is not a definitive judgment on the accuracy because certain predictions in the AD might be inaccurate, and vice versa (Sahigara et al. 2014). Similarly, Figure 9 shows that moderate and strong predictions are generally associated with higher concordance than are weak and very weak predictions.
Analysis and Interpretation of Metabolite Activity

The total number of computationally generated metabolites for ToxCast™ parent chemicals was 15,406 structures. This number includes simulations for first- and second-phase metabolism using the in silico tool ChemAxon Metabolizer (ChemAxon, Ltd.). The metabolite structures were standardized and deduplicated using the QSAR-ready workflow, producing 8,601 unique structures that partially overlapped with the prediction set (a total of 2,009 structures). The number of active and inactive metabolites for binding, agonist, and antagonist modalities are summarized in Table 9. These deduplicated, standardized metabolite structures, however, are generated from different parent ToxCast™ chemicals. The active metabolite structures were mapped back to their parent structures; i.e., the parent of each metabolite was identified. The major concern with metabolites is in situations when they are structures; i.e., the parent of each metabolite was identified. Table 10 summarizes the number of active metabolite structures predicted to be active using the consensus models.

Table 10. Active and inactive ToxCast™ parent chemicals linked to the metabolites predicted to be actives.

<table>
<thead>
<tr>
<th>Number of Active parent</th>
<th>Inactive parent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binding</td>
<td>145</td>
</tr>
<tr>
<td>Agonist</td>
<td>21</td>
</tr>
<tr>
<td>Antagonist</td>
<td>132</td>
</tr>
<tr>
<td>Total</td>
<td>152</td>
</tr>
</tbody>
</table>

Table 10 reveals that the number of inactive ToxCast™ parent chemicals with active metabolites is higher than those that were initially active according to the AR pathway model (based on ToxCast™ in vitro assays) (Kleinstreuer et al. 2017). The 212 inactive ToxCast™ chemicals with predicted active metabolites are likely candidates for follow-up as future work, either through experimental testing of the predicted metabolites, or through metabolically competent in vitro AR assays.

Generalization of the Consensus and Implementation in OPERA

To extend the use of CoMPARA to new chemicals by implementing support in the open-source prediction tool OPERA, a weighted kNN modeling approach was applied to provide predictions based on the existing experimental data and the prediction set consensus predictions with high concordance. Binding, agonist, and antagonist models were processed separately. To ensure high sensitivity for a conservative model, all actives were included, whereas inactives were set to a threshold of at least 85% concordance. ToxCast™-generated metabolite structures were excluded from this procedure (no means to validate against real metabolites structures). The total number of actives and inactives considered for modeling of each modality are reported in Table 11. Each of the three data sets was semi-randomly (stratified splitting) divided into training and test sets, each representing 75% and 25% of the actives and inactives, respectively. PaDEL and CDK2 descriptors were combined as described above. The GA-kNN procedure was conducted on the list of descriptors that passed the low variance and missing values filters. This step was conducted to make a supervised, end-point–dependent, similarity-based approach for the selection of the nearest neighbors.

The lists of chemicals summarized in Table 11 are provided as a single file containing identifiers, structures in SMILES format, and data in Supplemental Material S10, and as separate SDF files at https://doi.org/10.23645/epacomptox.10322012.

The statistics of the best kNN models, with k equal to five, for the three data sets are reported in Table 12. Because the goal of this modeling step was to reproduce the original consensus, the models were fitted, aiming for high performance and fidelity to the original predictions. However, the complexity of the models was kept to a minimum by adopting the weighted kNN approach and reducing the number of included descriptors. Indeed, for each modality (binding, agonist, and antagonist), the number of selected descriptors was optimized for minimum complexity and highest accuracy based on the GA ranking of importance. Figure 10 shows the selected number of descriptors for each model based on the ranked descriptors based on the GA results. In Figure 10, it is important to note that the ranking of the descriptors for the three modalities was not the same, because the three
GA procedures were conducted independently on the three data sets. Therefore, the three scatterplots are combined on the same graph for simplification and comparison reasons only.

Table 12 shows not only high BA but also good balance between Sn and Sp in five-fold cross-validation for the training set as well as the test set. The highest difference between Sp and Sn observed for the agonist model is probably the low number of actives (Table 11) in comparison with the binding and antagonist data sets. These statistics show that the three models are robust enough to simulate the original combined predictions and generate the same predictions for a new data set without having to rerun all the single models and repeating the whole procedure. Thus, the hypothesis of extending the consensus models to apply to new chemicals is valid.

These three weighted kNN models were first implemented in the OPERA (version 2.0) application, which is available for download from GitHub as command-line or user-friendly graphical interface (https://github.com/NIEHS/OPERA) (Mansouri et al. 2018). This application allows a user to generate CoMPARA (binding, agonist, and antagonist) predictions starting from a QSAR-ready 2D structure. This implementation of the models in OPERA will generate the same results as the initial combined predictions of the single models if tested on the same list of structures associated with the CoMPARA analysis. However, original predictions with concordance below the selected thresholds might change because they are not included in the knowledge base of the models.

Both OPERA’s AD indices as well as the nearest neighbors and the statistics are provided in the report that can be viewed in the application output or online on the EPA’s CompTox Chemicals Dashboard, and other OPERA models using the online CompTox prediction page (https://comptox.epa.gov/dashboard/predictions/index) or perform the calculations locally by installing the desktop standalone version (https://github.com/NIEHS/OPERA).

Figure 10. Selected descriptors for the binding (+ symbol), agonist (× symbol), and antagonist (♦ symbol) models and corresponding balanced accuracy (BA) calculated in five-fold cross-validation in forward selection based on the genetic algorithm (GA) ranking. The ranked descriptors are not overlapping for the three modalities.

Table 12. Training and test set statistics of the three consensus models.

<table>
<thead>
<tr>
<th>Number of Descriptors</th>
<th>Training set (75%) BA</th>
<th>Test set (25%) BA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sn</td>
<td>Sp</td>
</tr>
<tr>
<td>Binding</td>
<td>23</td>
<td>0.92</td>
</tr>
<tr>
<td>Agonist</td>
<td>10</td>
<td>0.92</td>
</tr>
<tr>
<td>Antagonist</td>
<td>15</td>
<td>0.91</td>
</tr>
</tbody>
</table>

Conclusions

The collaborative efforts of the CoMPARA participants resulted in robust consensus models predicting the potential ability of chemicals to interact with the AR pathway based only on their structures. Up to 91 separately developed categorical and continuous models were received from 25 international research groups. Separate models were built for agonist, antagonist, and binding activity based on a wide range of structure-based modeling approaches. The models were applied to a large collection of 55,450 chemical structures that included the 32,464 CERAPP chemicals as well as additional nonoverlapping chemicals from the European EINECS list and computational metabolites from ToxCast. CERAPP models were adapted, improved, and then applied at various stages of the project, from data and chemical structure curation to the evaluation of the submitted predictions and the consensus modeling. Most of the models were trained on a data set derived from the combination of 11 in vitro assays from ToxCast™ probing various points of the AR pathway model. Models were then evaluated using a collection of AR in vitro data curated from the online literature (PubChem). After this process, the categorical predictions were combined into consensus models to classify the chemicals as actives and inactives.

Despite the challenges caused by the skewed data distribution in terms of actives and inactives, most models achieved reasonably high performance, with a slight improvement for certain models.

Applications

This project has produced three generalized consensus models based on the initially generated predictions for the full CoMPARA list of 55,450 chemicals. A similar approach was applied to the CERAPP models that have also been implemented into the OPERA application, allowing them to be applied to new chemicals, reproducing the consensus predictions with high accuracy. The binding, agonist, and antagonist ER and AR activity models have been applied to the QSAR-ready forms of the entire set of 765,000 chemical substances contained in the EPA’s DSSTox database (underpinning the Dashboard application), and these predictions will be made available in the future via the EPA’s CompTox Chemicals Dashboard (https://comptox.epa.gov/dashboard) and the NTP’s Integrated Chemical Environment (ICE) dashboard (https://ice.ntp.nih.gov/). As for previously published OPERA models, once available on the dashboards, it will be possible to view the predictions for single chemicals online or downloaded using the batch search page (https://comptox.epa.gov/dashboard/dsstoxdb/batch_search). The online versions of the prediction reports will also include useful details, such as accuracy estimates and AD assessment. The details of the whole modeling procedure will be made available in a Qsar Model Report Format (QMRF) report that can be downloaded from OPERA’s prediction report page on the EPA Chemicals Dashboard or from the European Commission’s Joint Research Center (JRC) QMRF Inventory (European Commission 2013; JRC 2017). Moreover, the original online models developed by the VCCLAB team are available, together with all data, at http://ochem.eu/article/102271.

In addition to the precalculated predictions available on the CompTox and ICE dashboards for the full list of DSSTox chemicals, users will be able to perform real-time predictions for chemicals not contained in the database using the CERAPP, CoMPARA, and other OPERA models using the online CompTox prediction page (https://comptox.epa.gov/dashboard/predictions/index) or perform the calculations locally by installing the desktop standalone version (https://github.com/NIEHS/OPERA).
with narrow ADs. This achievement proves that there is not an optimal modeling approach for predicting these AR data. It is also important to note that, although the scoring function was used to combine the predictions from the different models, the concordance (majority rule) was the real driver of the final consensus predictions. Hence, most of the single models carried similar weights for an equal contribution to the consensus. The main difference between the models was their coverage of different portions of the prediction as determined by their defined AD, which also explains the fact that the concordance of the models with the final predictions on the full set of 55,450 chemicals (Figure 7) is different from the scores that were generated on the evaluation set, which represents only a small fraction.

The concordance among the predictions was high for most chemicals, particularly with inactives and strong actives. This consistency was demonstrated, as in the previous collaborative project CERAPP, to be linked to highly accurate predictions. This observation can therefore be extrapolated to new predictions. Low accuracy and concordance was noticed for weak actives, which, similar to ER activity, can be linked to the experimental differences between the ToxCast™-based training set and the literature-based evaluation. Relevant factors include but are not limited to lack of orthogonal assay results, differing concentration ranges, the presence of selective AR modulators (SARMs) with varying activity among tested tissue sources, and other experimental artifacts and errors.

The ultimate goal of this collaborative effort was to leverage the strengths of different modeling approaches in order to virtually screen a large universe of chemicals of high importance to environmental and human health studies. The final consensus models were able to identify approximately 10% of the screened chemicals as potential binders to the AR or in agonist/antagonist modes. This list included a number of computationally simulated ToxCast™ metabolites of inactive parent structures that require further in-depth attention. The resulting models were generalized and implemented in an open-source standalone application to be applied beyond the original list of screened chemicals. All materials and resulting files generated during this project are available for download at (https://doi.org/10.23645/epacomptox.10321697; https://doi.org/10.23645/epacomptox.10321982; https://doi.org/10.23645/epacomptox.10321925; https://doi.org/10.23645/epacomptox.10321994; https://doi.org/10.23645/epacomptox.10322012).

After CERAPP, which established the framework for such international collaborations, CoMPARA was a more global collaboration with a higher number of participants and a larger set of chemicals screened. The success of these two projects and the eagerness of the participants for more collaborations have prompted the organization of new consortiums to model new end points with readily available high-quality data, such as acute oral systemic toxicity (Kleinsteuber et al. 2018b). In summary, this project further demonstrates the power of combined computational approaches to accurately and rapidly screen large libraries of chemicals with high toxicological relevance and to provide open-source tools that can be readily applied by a wide range of stakeholders.

Acknowledgments

This work was supported by Oak Ridge Institute for Science and Education (ORISE) Research Participation Program at the EPA, the Lush Prize (Young Researcher) (2017), and The Intramural Research Program of National Institute of Environmental Health Sciences (NIEHS). Technical support was provided by Integrated Laboratory Systems Inc. under NIEHS contract HHSN273201500010C. S.S., G.P., A.T.G.S. and U.M. from University of Tartu are grateful for support from the Ministry of Education and Research, Republic of Estonia (grant number IUT34-14) and the European Union European Regional Development Fund through Foundation Archimedes (grant number TK143, Centre of Excellence in Molecular Cell Engineering).

The views expressed in this article are those of the authors and do not necessarily reflect the views of policies of the EPA or any other agency and should not be construed to represent any agency determination or policy. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

References


