Supplementary information: sAOP: Linking chemical stressors to Adverse Outcomes Pathway Networks.

Alejandro Aguayo-Orozco, Karine Audouze, Troels Siggaard, Robert Barouki, Søren Brunak, Olivier Taboureau.

Materials and methods

Chemical and biological targets data

Relevant information regarding links between chemical substances and biological information was collected using the latest update of the ToxCast database (as of March 2018). The ToxCast database contains information for 9,076 chemicals that have been tested in a total of 359 assays endpoints corresponding to a total of 1,192 endpoints (e.g. two endpoints can be the results from the same assays endpoints but measured at different times). Endpoints can measure changes at molecular level, such as protein conformation and DNA binding, and changes in cell behaviour, such as organelle disruption or cell viability. Details regarding the different platforms, technologies and assay endpoints have been described previously (Judson et al., 2016) and can be found on the interactive Chemical Safety and Sustainability ToxCast dashboard (https://actor.epa.gov/dashboard; last accessed March 5, 2018). For the present study, the ToxCast database was downloaded as a Structured Query Language (SQL) database, which is a domain-specific language used in programming and managing of data based on a relational database management system. It included information on assay endpoints, chemicals, targets and their activity resulting from the high-throughput analyses. For each assay endpoint, all relevant information regarding type of assay endpoints, protocols, the input, output and controls are mentioned. Moreover, the chemical compounds are listed with name, CAS number and the assays tested. For each assay endpoint, there is an intended target (ranged from molecular interactions to cellular organization and viability). In many cases, chemicals have been tested at several concentrations for an assay endpoint allowing fitting the corresponding dose-response curve. For each chemical-assay studied, information regarding AC50 (concentration at 50% of the maximum activity calculated through Hill or gain/loss methods) reported in the database was considered in our analysis.

Data curation/pre-processing

Preparation of data from AOP wiki

The AOP-Wiki database was considered as a source of information for AOPs data (https://aopwiki.org). The current version (as of March 2018) allowed to extract 207 AOPs. In order to facilitate further connection between each AOP and ToxCast database, it is necessary

to describe the content of the AOPs. 1,105 KEs were extracted from the application programming interface (API) in a tabulated format, implemented within AOP wiki. MIEs, which are specialized KEs that describe the initial interaction between a stressor and a biomolecule, can involve specific proteins and genes such as Cyp19a1 or AchE. Other KEs can play the role of MIE in one AOP and the role of intermediate KE in another. This is the case for example of *Activation, Glucocorticoid Receptor*. Among all the AOPs' information downloaded, a total of 117 MIEs, 892 KEs, 96 AOs and 23 MIEs/KEs were kept for mapping. Furthermore, the data extracted from the AOPs were also annotated as 'Up' or 'Down' depending on the directionality of the event. The decision was made based on the description of the event (under the field "name"). Therefore, through natural language processing, the terms with synonymous definition were treated equally, such as *agonism, activation, increasing,* which point to a positive directionality of the event. The list of synonyms used is available in supplementary information (Table S1) Among all KEs, 418 are events that move in the positive direction and, 298 in the negative direction and 389 that could not be defined with a specific directionality, hence considered to be affected in both directions, i.e. 'Altered regulation, alpha haemoglobin'. In order to prepare the data for the further mapping steps, the MIEs were manually annotated with the associated genes/proteins name when possible i.e. '*Agonism, Androgen Receptor'* was associated to the protein androgen receptor (AR). In other cases, when several isoforms were present, all isoforms were linked and annotated to the MIE. For example, the two isoforms of the LXR proteins (LXRα and LXRβ) were associated to '*Activation of LXR'*. Some MIEs were unspecific and could be associated with genes or proteins. This was the case of the MIE '*Activation, Hepatic nuclear receptor(s)'*, which was annotated with nuclear receptor genes or proteins existing in the nucleus of hepatic cells. Finally, to facilitate further analysis, UniProt identifiers were used for all the gene/protein associated in MIEs. For each gene/protein, the human uniprot identifier was considered.

Table S1: Synonyms used in the natural language processing step in methods. With these all KEs were mapped to Up or Down regulation.

Up	**Down**
Accelerated	Antagonism
Accumulation	Anticoagulant rodenticide interferes with gamma-carboxyglutamate formation
Activate	Binding as antagonist
Activated	Binding of antagonist
Activation of TGF-α signaling	Binding of inhibitor
Activtation	Changes/Inhibition
Activation of Cyp2E1 in the liver	Decline
Activation of specific nuclear receptors	Decrease
Activation of TGF-α signaling	Decreased
Activation	Delay
Activation/Proliferation	Delayed

Agonism	Depletion
Airway Hyper-responsiveness	Desensitization
Excitation	Direct mitochondrial inhibition
Binding of agonist	Down Regulation
Formation	Impaired recruitment
Increase activation	impaired
Increase proliferation	Impairment
Increase	Inactivated
Increased activity	Inactivation of PPARŒ≥
Increased Apoptosis	Irreversible inhibition of hepatic VKOR by binding of AR at tyrosine 139
Increased CGRP	Inactive
increased mantel display	Inhibit
Increased NKA	Inhibition of Ca Channels
increased or inappropriate	Inhibition of neurotransmitter release
Increased	Inhibition
Induced parturition	interruption
induced spawning	Mitochondrial impairment
induced	Reduce expression
Induction	reduced dimerization
Mu Opioid Receptor Agonism	Reduced fitness or even mortality
Opening of GIRK channels	reduced production
Opening of calcium channel	Reduced
Overactivation	reduction in ovarian granulosa cells
prepubertal increase	Reduction
Pro-inflammatory cytokines increased	Suppression
Production of Œ±-smooth muscle actin	Weakened
Production	
Proliferation	
Proliferation/Clonal Expansion	
prolonged	
Propagation	
Serotonin 1A Receptor Agonism	
Synthesis	
Systemic cholesterol increased	
TH synthesis	
TRPA1 activation	
TRPV1 activation	
Tumorigenesis	
Up Regulation	
Upregulated	

Preparation of data from ToxCast

In our implementation, the association between assay and protein is based on the ToxCast definition, i.e. in many case one assay is related to one protein or gene. There are more complex relationships. For some assays, this is not only one protein but a set of proteins that are analyzed. It means that we do not know if a chemical has an effect on all the proteins or only one. In this case, ToxCast considered that the entire set of proteins is targeted by the chemical. We used this annotation as well in our mapping.

From the ToxCast SQL database, the AC50 in logarithmic values (logAC50) were extracted for the all experiments chemical-assay endpoint. Not all the compounds have been tested against all the assay endpoints, hence all the data points with missing values were removed for further analysis. After this first filtering step, a total of 2,386,156 data points, containing the chemical-assay endpoint tested data, were gathered. For some assay endpoints a single concentration was tested, hence not producing the necessary data points to produce an AC50 value, which is denoted with an AC50 value of 0 in the database. For this filtering step the chemical-assay activity, characterised by the hit calls, was used. If the data resulting from a chemical-assay endpoint pair fits one or more of the models, this pair is considered "active", which equals to an active "hit call". Hit call contains values of 1, 0 or -1, meaning active (1), inactive (0) or undetermined activity (-1) (Judson, 2015, 2016). The active ones were selected, hence removing all those that contain 0 or -1 in their hit calls (486,867 data points). Furthermore, in order to be considered toxic, a compound with AC50 different from 0 has to fit a Hill or Gain/Loss model, otherwise this compound was removed. To fit a Hill model, a chemical has to be tested at multiple concentrations. Consequently, those compounds that were not treated at multiple concentrations or did not fit the models where removed. This filter gave rise to 6,569 chemicals showing toxicity in at least one assay endpoint, and 974 endpoints that showed activity under the presence of at least one compound.

Z-score, which is an indicator of the distance from cytotoxicity distribution, was applied to assess the relationship between chemical-assay endpoint and assay endpoint cytotoxicity distributions as suggested by Judson et al. (Judson, 2016). According to Judson et al. high z-score chemical-assay endpoint hits occurred in concentration regions where there was no evidence of cytotoxicity or cell stress, hence they hypothesized that those hits were more likely to be associated with the intended biological process or target than to cytotoxic effects. The z-score (Eq. 1) calculation allowed removing chemicals that showed CTB at a lower concentration compared to the AC50 value.

Eq. 1

$$Z(\text{chemical, assay}) = -\text{logAC50(chemical, assay)} - \text{median}[-\text{logAC50(chemical,cytotox)}] / \text{global cytotoxicity median absolute deviation}$$

In this study, we used a global cytotoxicity median absolute deviation of 0,293 based on 33 cytotoxicity assays, as defined by Judson et al. A chemical with a z-score, in absolute value, higher or equal to 2.0 was considered non-cytotoxic as reported by Auerback et al. (Auerback et al. 2016). A total of 82,708 data points (chemical-assays endpoints combination) fitted in this threshold of cytotoxicity (value below 2.0) and have been excluded from further analyses. It should be noted that a z-score cut-off of 3 was also proposed by Judson et al. (Judson et al., 2016). The user can select this more stringent cut-off on the sAOP tool.

From this analysis, a total of 6,111 chemical compounds showed non-cytotoxic activity in at least one assay endpoint, and 906 endpoints targeted by at least one non-cytotoxic compound were conserved. A workflow of the proposed strategy for the curation and the mapping of adverse outcomes to chemical substances is shown in Figure S1.

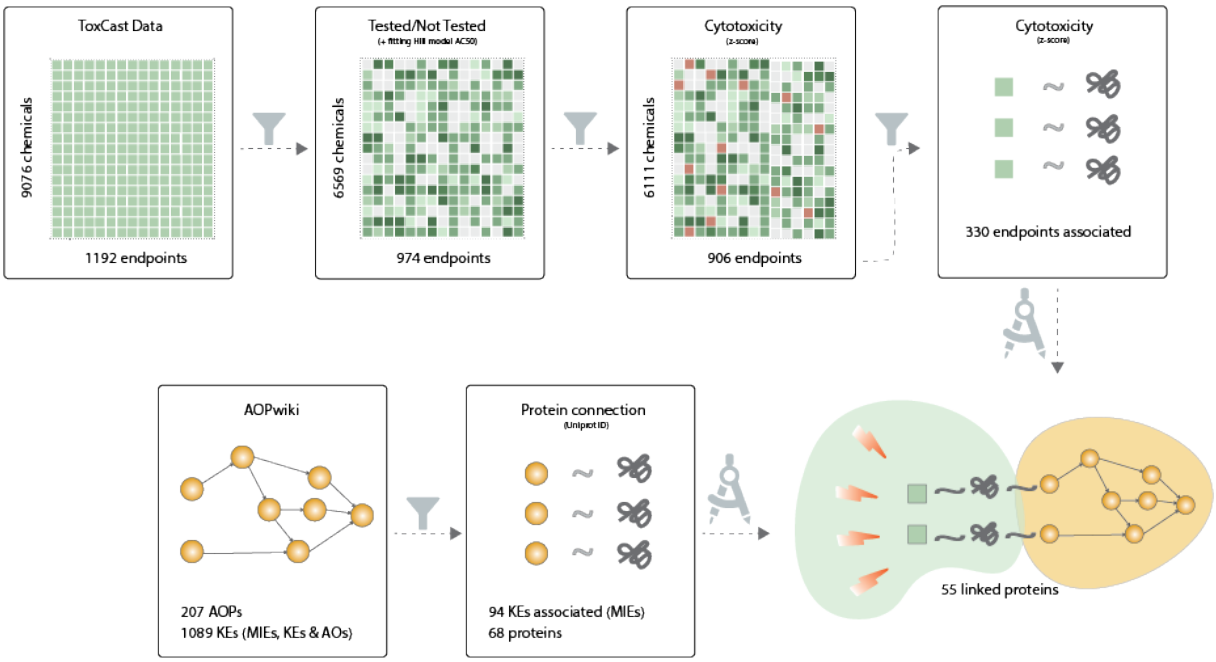

Figure S1. General framework for mapping 207 AOPs to 906 assay endpoints and thereby to 6,111 chemicals. The workflow indicates the filtering steps required to obtain the endpoints and compounds that will map with the Key Events (KEs) in AOPs that correspond to proteins. 1A) represents the entire ToxCast chemical-assay endpoint data space where no filtering has been yet applied. After the first filtering step in 1B), the hit calls with value 1 have been selected and those results that fit a Hill or Gain/Loss model. The different tones of green in the image represent the differences in AC50. 1C) Next filtering step is based on cytotoxicity, in which the red spots represent the chemical-assay combinations that have a z-score below 2.0. The last step in 1D) associate an UniProt ID to an endpoint (the green squares represent the endpoint and the grey figure represents the protein). On the below left representation, the KEs are represented by yellow nodes, which are in the next image associated to Uniprot IDs. The last image represents the link between assays and KEs by the associated Uniprot ID, and how these are the interconnected with the rest of KEs on one side, and by the active chemicals on the other.

Matching bioassays from ToxCast and MIE (KE) from AOP

In ToxCast any protein or gene are annotated with a HUGO name. So, in AOP wiki, all MIEs and KEs which are defined by a protein or a gene were annotated also with a HUGO name in

order to facilitate the connection between ToxCast and AOP wiki. From a total number of 1192 of endpoints from ToxCast, 906 endpoints representing 369 unique proteins can be connected to one or more of the 207 different AOPs extracted from the AOP-Wiki. Mapping, this set of protein from ToxCast to the list of KEs from AOP-Wiki, we found 68 proteins that are described in 94 KEs. However, 14 MIEs associated to proteins, could not be mapped to ToxCast due to a lack of assay that measures the activity in the adequate direction. For example, KE 471 (Inhibition, FoxA2) was related to the assay endpoint, ATG_FoxA2_CIS_up. But, this assay measured the agonist activity. Hence, this assay was not used to explain KE 471. Removing the list of proteins for which no assays with adequate direction (Q9Y261, P05181, Q9Y243, Q92851, P45984, P28482, Q15759, P27361, Q8IW41, Q96EB6, P28223, P07550, P28335 (based on UniProt ID)), we finally obtained 55 proteins (from 126 endpoints) targeted by 4960 chemicals and matching with 94 KEs. The lower number of unique different proteins outcome in ToxCast is due to the redundancy of these proteins in different assays in ToxCast.

We have noticed that a link from a gene to an AOP-Wiki is also available in the dashboard of ToxCast. Still the sAOP tool is complementary as our tool allows to visualize the connection in an AOP network and to suggest new hypotheses of associations between chemicals, proteins, MIEs, KEs and AOs.

sAOP web application

A web application was designed to visualize and extract (stressor) adverse outcome pathways (AOP), chemicals, proteins, assays, key events, adverse outcomes and their interactions. Software development was carried out with popular programming languages and architectural methods such as Javascript, Apache server with a MySQL database, PHP, a RESTful API, Cytoscape.js for network visualization.
Querying the database request parameters: ‘search by’, ‘search for’, ‘degree of separation’, ‘z-score’ and ‘AC50 score’ which are sent to the RESTful API translating the data to a SQL query and returns a result in JSON format. The JSON formatted data can be downloaded as a file, for further analysis e.g. in Cytoscape for Desktop. For TSV-download functionality, the JSON data is converted into tab separated values and presented as two downloadable files, nodes and edges. The visualization page also offers the ability to generate a JPG image file for download. User guidelines for the web application can be found also in the following link:

http://saop.cpr.ku.dk/?page=help.

The degree of separation allows to select the number of connections away from a query. For example, query for a chemical with a degree of separation of 1, will provide a network with the bioassays on which the chemical is bioactive. A degree of separation of 2 will show the bioassays and the proteins. A degree of separation of 3 from a chemical will depict in addition MIEs and/or KEs and it start to be possible to detect some AOs with a degree of separation of 4 from a chemical. So, by increasing the degree of separation, the network will increment in complexity. If a user starts from a protein or a KE and use a degree of separation of 1, the network will represent the first link with any other sources upstream and downstream of the query.

How the chemical stressor-AOP network can be used?

Integration of the AOPs information with the high throughput data from the ToxCast database

can help 1) to provide a list of potential chemical stressors directly associated to an AOP and 2) to propose new links between MIEs and KEs leading to an AO not already reported in AOP-Wiki. Two examples are described below to demonstrate the relevance of the sAOP approach in the assessment of chemical stressors-AOP and for the discovery of new MIEs and KEs associated to AOPs.

Suggestion of new MIEs and KEs associated to AOs

During the integration of the data, we observed that some KEs can be connected to different AOs, although they are only described for one AOP. This is the case for example with the AOP-Wiki ID 30, named 'Estrogen receptor antagonism leading to reproductive dysfunction', for which, one MIE (Antagonism estrogen receptor), 4 KEs (Vitellogenin synthesis in liver, Plasma vitellogenin concentrations, Vitellogenin accumulation into oocytes and oocyte growth/development, Cumulative fecundity and spawning) and one AO (population in trajectory) are associated. Through the sAOP network, other KEs which are not part of the defined AOP are affected downstream the MIE and can lead to toxicity through other AOPs. For example, the KE 25 (Agonism, androgen receptor and Reduction) is associated with two others AOPs (AOP 7: Aromatase inhibition leading to ovulation inhibition and decreased fertility in female rats, and AOP 23: Androgen receptor agonism leading to reproductive dysfunction (in repeat-spawning fish).

Additionally, other chemicals can also be linked to the AOPs through the proteins they affect and the related MIEs and KEs. This leads to downstream events in various AOPs, contributing information about chemicals which affect those KEs and eventually produce AOs (Figure S2).

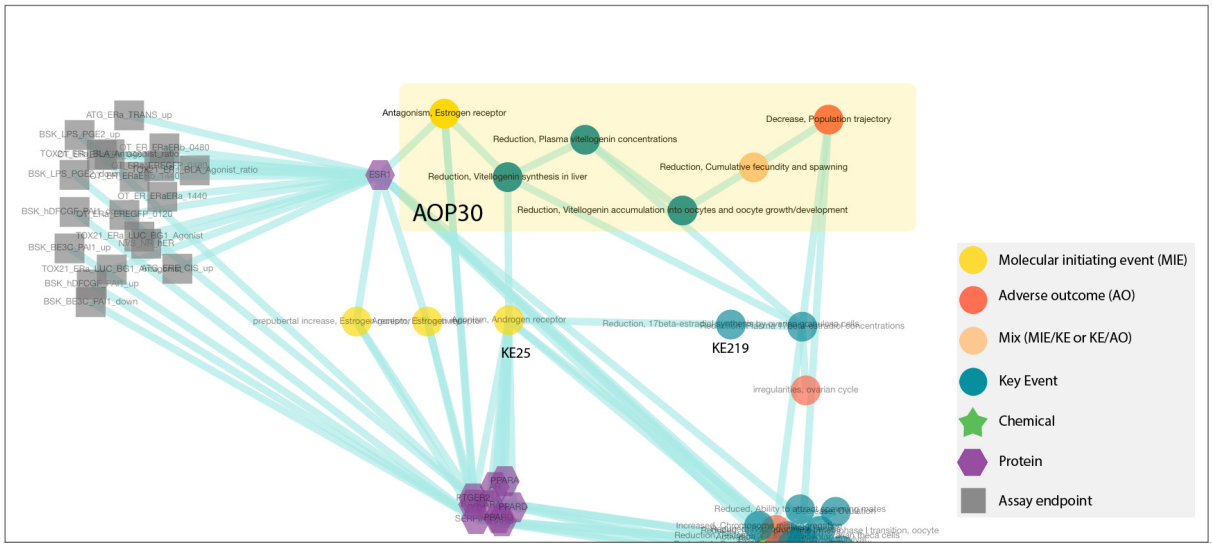

Figure S2. Snapshot of the web application sAOP output network. The protein estrogen receptor 1 (ESR1) (in violet) is involved in the MIE Antagonism Estrogen receptor of the AOP30 which contain 4 KEs (green and light orange sphere) and one AO (red sphere). It plays also a role in other AOPs due to its connections to other KEs that do not belong to AOP 30. The ESR1 receptor is connected to several assay endpoints (grey squares) which are associated to other proteins affecting other KEs. The yellow box highlights the AOP 30. The active chemicals on the assay endpoints are not depicted in this figure.

Linking AOPs to a single substance: BPA

Bisphenol A (BPA) is a suspected EDC, which may be related to many toxicological effects on human health. From the developed AOP network, BPA is active on 120 proteins in ToxCast via 19 assay endpoints, which have been mapped to 22 KEs. Not surprisingly, many of the KEs are related to androgen and estrogen activities, as well as to other nuclear receptors (NR1H4, AhR, PXR, LXR, PPARg, GR). Therefore, BPA may affect several biological pathways involved in various human diseases such as metabolic disorders. An interesting result is the association between BPA and obesity, involving the MIE, PPAR-gamma activation in hepatocytes (AOP-Wiki ID: 1028) and two KEs, activation CEBPA (AOP-Wiki ID 1448) and increase adipogenesis (AOP-Wiki ID 1449) leading to obesity (AOP-Wiki ID 1447) (Figure S3). In ToxCast, BPA activates PPARg in human liver cells (HepG2). Such activation is directly linked to promote adipogenesis, which is a causative factor for obesity (Legeay, 2017). Regarding BPA, growing evidence includes epidemiological data, in vivo, in vitro and computational studies indicating that BPA is positively correlated with an increased risk of obesity and should therefore be considered as an obesogenic compound. It has been demonstrated that BPA exposure increases the body weight and mass index (Junge, 2018). The findings obtained here reinforce the potential contribution of BPA to the pathophysiology of obesity.

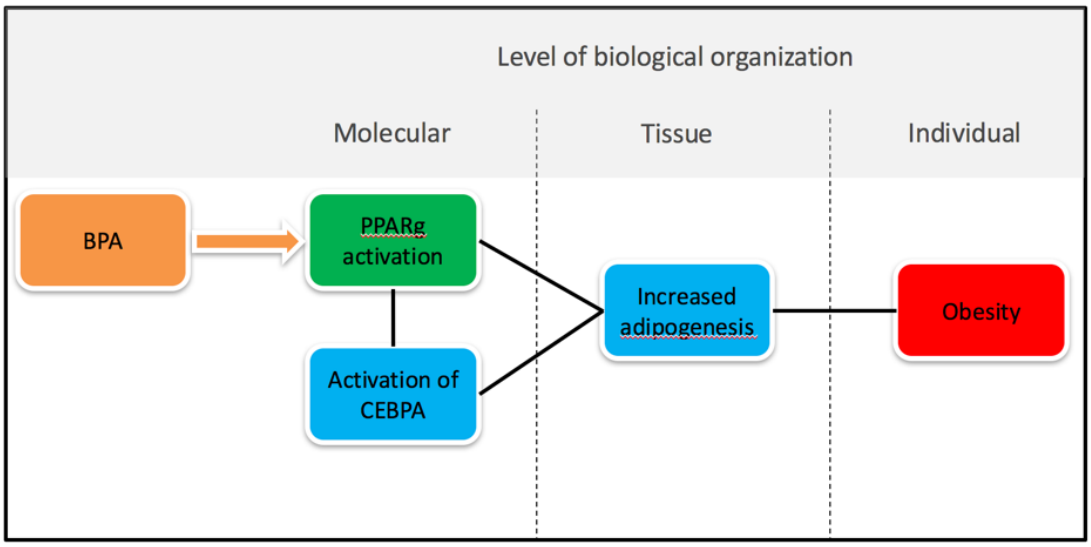

Figure S3. Linking bisphenol A (BPA) to obesity. The stressor-AOP network allows suggestion of a potential mechanistic linkage between BPA and obesity. The initial step (in orange) represents BPA as a potential stressor. The activation of PPARγ by the stressor defines the MIE (in green), which will induce two key events (in blue) at different levels of the biological organization (as defined in the AOP-Wiki database). The AO (in red) represents the clinical effect at the individual level.

Notification

This knowledge-based network has some limitations and is clearly dependent on the available data. The 207 AOPs described in the AOP-Wiki database are at various stages of development and not all of them are approved by OECD. Although a systematic collection of weights of evidence is implemented (Hoffmann & Hartung, 2006; Hoffman et al., 2017), several AOPs

have zero weight of evidence or key event description and feedback from the scientific community is important for the quality control of AOPs in the AOP-Wiki (Lalone and Hecker, 2017). It is expected that the extension of reviewed AOPs with additional weight of evidence will enrich our stressor-AOP network. Furthermore, in this first version of the sAOP, only human proteins and AOP developed for human species have been considered. We will plan to include other species in the next version of the sAOP.

References

Auerback, S.S. et al. (2016). Prioritizing environmental chemicals for obesity and diabetes outcomes research: A screening approach using ToxCast high-throughput data. Environ. Health Perspect. 124, 1141-54

Hoffmann, S. et al. (2017) A primer on systematic review in toxicology. Arch Toxicol., 91, 2551-2575.

Hoffmann, S. & Hartung, T. (2006) Towards an evidence-based toxicology. Hum. Exp. Toxicol., 25, 497-513.

Judson, R.S. et al. (2015) Integrated Model of Chemical Perturbations of a Biological Pathway Using 18 In Vitro High-Throughput Screening Assays for the Estrogen Receptor. *Toxicol. Sci*, **148**, 137-154.

Judson, R.S. et al. (2016) Analysis of the effects of cell stress and cytotoxicity on in vitro assay activity across a diverse chemical and assay space. *Toxicol. Sci*., **152**, 323-339.

Junge, K.M. et al. (2018) MEST mediates the impact of prenatal bisphenol A exposure on long-term body weight development. *Clin. Epigenetics*, **10**, 58.

LaLone, C.A. et al. (2017) Society of Environmental Toxicology and Chemistry (SETAC) Pellston Workshop: Advancing the Adverse Outcome Pathway Concept: An International Horizon Scanning Approach, April 2-6, 2017, Cornwall, ON, Canada. Society for the Advancement of AOPs, Durham, North Carolina, USA. [cited 2017 August 8]. Available from: http://www.saaop.org/workshops/pellston2017.html

Legeay, S. & Faure, S. (2017) Is bisphenol A an environmental obesogen? *Fundam. Clin. Pharmacol*. **31**, 594-609.