Timothy E. H. Allen Andrew J. Wedlake Fiona Goldhoyre Charles Gong Jonathan M. Goodman Steve Gutsell Paul J. Russell

Assessing Machine Learning Methods in the Identification and Quantification of Environmental Chemical-Key Event Pairs Associated with Adverse Health Outcomes

Holly M. Mortensen[A], Timothy E.H. Allen[B], Jonathan Senn[C], Grace Patlewicz[D], Imran Shah[D]

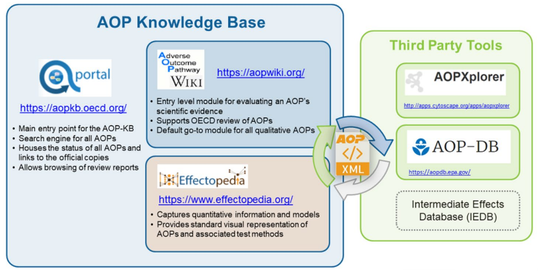
[A]US EPA, ORD, CPHEA, RTP, NC USA; [B]MRC Toxicology Unit, University of Cambridge, UK; [C] Oak Ridge Associated Universities, RTP, NC USA; [D] US EPA, ORD, CCTE, RTP, NC USA.

Holly M. Mortensen mortensen.holly@epa.gov
Poster #P105 Abstract #2367

Background

The ability to link the chemistry of chemical toxicants to adverse outcomes at high levels of biological organization allows for greater understanding of the biological mechanisms responsible for toxicity. The adverse outcome pathway (AOP) construct is useful in establishing and documenting the biological mechanism(s) implicated in adverse health outcomes of toxicological concern. Though the AOP construct is chemically agnostic by definition, there is an interest in identifying chemical groups that are known to affect particular pathway function, specifically in the identification of molecular initiating events (MIEs) that are both critical in the progression of the outcome and potential candidates for *in vitro* assay development. Modelling these MIEs and linking them to higher level key events allows for the effective combination of *in vitro* and *in silico* toxicology tools, contributing to new approach methodologies (NAMs) for use in safety evaluation.

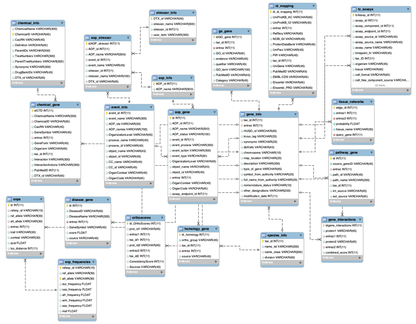
AOP-DB v.2



Currently, the AOP-DB v.2 stores chemical stressor information from the AOP-Wiki XML, which are derived from expert submissions to the AOPWiki, but cannot be depended on for accuracy, consistency, or completeness given the crowd-sourced nature of the AOPWiki.

AOP-DB v.2 introduces several substantial updates, which help inform new testable hypotheses about the etiology and mechanisms underlying adverse outcomes of environmental and toxicological concern[1,2,3].

One area of interest is the computational estimation of association between a chemical stressor and a molecular target. Because AOPs are, by definition, chemically agnostic, there has been disagreement on how to interpret and quantify the relationship between chemical stressor and key events associated with AOPs.



Dataset Construction and MIE Predictions

Here we aim to extend the machine learning methodology from the molecular to the adverse outcome, in order to increase confidence and provide mechanistic information on the chemical. In this way we can provide a "prototype stressor" for adverse outcomes of toxicological interest.

In an effort to identify prototype stressors for AOPs, and quantify the level of association between stressor and MIE (e.g. causal stressor-target relationships) we have:

- Used SQL to query the AOP-DB v.2, and generate an output, including aop_name, event_id, event_name, AOP_id, stressor_name, and HUGO identifier.
- Removed cases where the stressor was listed as "NULL" or "n/a" then duplicate stressors.
- Restricted the output to KEs that had previously constructed machine learning models[4].
- Generated predictions for each SMILES string in the NN models described[4], including a similarity search for the most similar training set chemical.
- Conducted a search of the ChEMBL database (https://www.ebi.ac.uk/chembl/) to find experimental activity values (pChEMBL values) for these chemicals at their expected MIEs.

AOP ID	Stressor Name	Target Gene ID	NN Prediction	AD	pChEMBL Value	Note
112	Bromocriptine	ER	0.118	1.000	Not active	Dataset Inactive
14	Cinnamic aldehyde	NR3C1	0.131	1.000	Not active	Dataset Inactive
111	Vinclozalin	AR	0.166	1.000	4.35	Dataset Inactive
28	Aceytlsalicylic acid	PTGS1	0.109	1.000	4.00-6.52	Dataset Borderline
117	Androstenedione	AR	0.406	1.000	4.75-6.65	Dataset Borderline
28	Celecoxib	PTGS1	0.373	1.000	4.09-6.68	Dataset Borderline
28	Ibuprofen	PTGS1	0.643	1.000	4.14-5.97	Dataset Borderline
28	Naproxen	PTGS1	0.653	1.000	4.28-6.75	Dataset Borderline
23	5α-Dihydrotestosterone	AR	0.851	1.000	7.88-9.70	Dataset Active
14	Dexamethasone	NR3C1	0.943	1.000	7.68-9.30	Dataset Active
28	Diclofenac sodium	PTGS1	0.432	1.000	5.20-7.17	Dataset Active
25	Fadrozole	CYP19A1	0.946	1.000	6.70-10.30	Dataset Active
97	Fluoxetine	SLC6A4	0.966	1.000	6.75-9.14	Dataset Active
111	Flutamide	AR	0.316	1.000	6.24-6.81	Dataset Active
28	Indomethacin	PTGS1	0.818	1.000	5.05-8.40	Dataset Active
25	Letrozole	CYP19A1	0.934	1.000	7.30-10.70	Dataset Active
43	Sunitinib malate	VEGFR-1	0.905	1.000	8.70-9.00	Dataset Active
43	Vatalanib	VEGFR-1	0.965	1.000	6.42-8.02	Dataset Active
23	17beta-Trenbolone	AR	0.288	1.000	8.58 in rat	Dataset Active (rat)
23	Spironolactone	AR	0.558	0.312	6.17-7.41	Correct Prediction
34	Troglitazone	PPARG	0.517	0.459	5.42-6.52	Correct Prediction
36	Benzo(k)fluoranthene	PPARG	0.347	0.218	5.25-5.35	Incorrect Prediction

Table showing identified stressor-target-AOP relationships from AOP-DB v.2 along with NN predictions and applicability domain values (AD) for stressors with activity data at their MIE target in ChEMBL. Data from ChEMBL helps identify if a stressor is active (green), inactive (red) or borderline active (yellow) at the suspected molecular target.

Stressor-Target-AOP Discussion

This project links 79 machine learning neural network models constructed using publicly available data for the prediction of MIEs to 280 unique AOPs (1111 KEs and 193 MIEs) in the AOP-DB v.2. Neural network (NN) model performance has previously been evaluated and prediction accuracy is above 90% on test and 75% on external validation data.

Of the 22 stressor-target-AOP relationships from AOP-DB v.2 with experimental data in ChEMBL, 19 were found to be in the training data used to construct the NN models, with 8 being listed as inactive. Three chemicals were identified as inactive at their supposed target and five as borderline active (with pChEMBL = 5 used as the active/inactive threshold) in the ChEMBL Database. Additionally, of the 15 stressors with no experimental data in ChEMBL 13 have been predicted as inactive at their MIEs by the NN models.

Of these stressor-target relationships, bromocriptine binding to the estrogen receptor, cinnamic aldehyde binding to the glucocorticoid receptor and vinclozalin binding to the androgen receptor are most unlikely to be responsible for any measured adverse outcomes through their AOPs.

These finding suggest further investigations into the toxicity mechanisms of these chemicals are appropriate to properly understand their AOPs.

Of the four cases where the stressor was not included in the NN training set, one contained no appropriate data in ChEMBL, two were correctly predicted by the machine learning algorithms as active and one was incorrectly classified as inactive.

Future Directions

The combined ChEMBL/ToxCast information used to build the models[4] and implemented here, suggest the molecules indicated in red are not binders at the MIE claimed in the AOP-DB. Additional investigation into the toxicity mechanisms in the cases identified will be necessary. Additionally, improved annotation and data structure for stressor-key event information in the AOPWiki is upcoming, and may help clarify stressor-target-AOP relationships, and potential expansion on the number of MIEs that can be predicted.

References

1. Pittman, M.E., Edwards, S.W., Ives, C., Mortensen, H.M. (2018) AOP-DB: A database resource for the exploration of Adverse Outcome Pathways through integrated association networks. Toxicology and Applied Pharmacology Mar 15; 343:71-83. doi: 10.1016/j.taap.2018.02.006. Epub 2018 Feb 14.
2. Mortensen, H.M., Chamberlin, J., Joubert, B., Angrish, M, Sipes, N., Lee, J.S., Euling, S.Y. *(2018)* Leveraging human genetic and adverse outcome pathway (AOP) data to inform susceptibility in human health risk assessment. *Mammalian Genome*. Feb;29(1-2):190-204. doi: 10.1007/s00335-018-9738-7. Epub 2018 Feb 23.
3. Mortensen, H.M., Levey, T., Langley, P., Senn, J., Williams, A. *(submitted)* EPA AOP-DB version 2 updates. *Nature Scientific Data.*
4. Allen, TEH, Wedlake, AJ., et al. *Chem. Sci.*, 2020,11, 7335-7348.

The views expressed in this poster are those of the authors and may not reflect U.S. EPA policy