Using automated machine learning for the prediction of developmental and reproductive toxicity

Showcase week (Theory)

Marcus Wang

26 Sep 2022

UNIVERSITY OF CAMBRIDGE

Developmental and reproductive toxicity (DART)

- Reproductive
  - Effects on fertility (male/female), childbirth, lactation
- Developmental
  - Related to the fetus
  - Mortality, structural alterations, growth, functional deficits

Developmental and reproductive toxicity (DART)

- Limited quality data
- Expensive to measure
- Long duration to get results
- Ethical issues
- Regulatory requirements

Data sources

- Data was compiled from a total of 12 different sources
  - includes both *in vitro* and *in vivo* data
  - Includes data from publicly available sources as well as commercial software eg. DEREK
  - Includes drug-like chemicals as well as industrial chemicals (ToxCast)
- Database includes data covering the endpoints:
  - Teratogenicity
  - hERG channel inhibition
  - Steroidogenesis oestrogen
  - Prenatal developmental toxicity
  - Sperm reduction, gonadal dysgenesis, abnormal ovulation, and infertility growth retardation

How the data is processed

- Toxicity value recorded for each compound by source
- Entries with missing or unclear data was removed
- Salts and metals were removed
- Database was checked for duplicates entries based on InChI Key, SMILES, CAS columns
- Enantiomers were merged
- Entries for the same compound were merged
- Overall toxicity value determined to be positive (1) if any of the tested sources recorded a positive

Database statistics

- 3255 compounds (no salts/metals)
  - 1672 positives
  - 1583 negatives
- Largest known database for DART that has been compiled

Automated machine learning (AutoML)

Automated machine learning (AutoML)

- Automating the machine learning process
- Optimise model hyperparameters automatically
- Automated feature selection/processing also possible

Automated machine learning (AutoML)

- More efficient than manually specifying and adjusting hyperparameters
- Increasing popularity in recent years

Automated machine learning (AutoML)

- AutoGluon package used
- Includes common machine learning models eg. Random forest, ANN

Model metrics

- Proportion of true positives

$$Sensitivity\ (SE) = \frac{TP}{TP + FN}$$

- Proportion of true negatives

$$Specificity\ (SP) = \frac{TN}{TN + FP}$$

Model metrics

- Correct classification by model

$$Accuracy\ (Q) = \frac{TN + TP}{TN + FP + TP + FN}$$

- Matthews correlation coefficient (MCC)

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Preparing the data for input

- Per run for a total of 5 runs

Database (ca. 3250 compounds)

↓ Random split

Remaining data (80%)	Test (20%)

↓ Random split

5-folds (64% training, 16% validation)

Benchmarking results

Model	SE (%)	SP (%)	Accuracy (%)	MCC
CatBoost_BAG_L1	77.9 ± 4.9	87.8 ± 1.8	83.1 ± 2.2	0.662 ± 0.044
CatBoost_BAG_L2	78.9 ± 5.0	90.7 ± 3.8	85.1 ± 3.2	0.704 ± 0.067
ExtraTreesEntr_BAG_L1	79.5 ± 4.6	89.7 ± 3.7	84.9 ± 3.2	0.698 ± 0.067
ExtraTreesEntr_BAG_L2	79.0 ± 4.9	90.8 ± 3.5	85.3 ± 2.8	0.707 ± 0.059
ExtraTreesGini_BAG_L1	79.9 ± 4.8	90.0 ± 3.4	85.2 ± 3.4	0.704 ± 0.070
ExtraTreesGini_BAG_L2	79.7 ± 4.7	89.8 ± 3.7	85.1 ± 3.3	0.702 ± 0.067
LightGBMLarge_BAG_L1	80.8 ± 3.6	84.9 ± 3.6	83.0 ± 2.9	0.659 ± 0.059
LightGBMLarge_BAG_L2	78.1 ± 4.2	88.6 ± 4.9	83.6 ± 3.4	0.673 ± 0.072
LightGBMXT_BAG_L1	81.5 ± 3.2	84.4 ± 3.5	83.1 ± 2.8	0.66 ± 0.058
LightGBMXT_BAG_L2	78.9 ± 4.6	90.9 ± 2.9	85.3 ± 2.7	0.707 ± 0.055
LightGBM_BAG_L1	81.5 ± 3.2	84.4 ± 3.5	83.1 ± 2.8	0.66 ± 0.058
LightGBM_BAG_L2	79.3 ± 5.2	90.0 ± 4.1	84.9 ± 3.2	0.70 ± 0.068
NeuralNetFastAI_BAG_L1	83.7 ± 5.1	79.4 ± 2.8	81.4 ± 2.3	0.63 ± 0.046
NeuralNetFastAI_BAG_L2	83.9 ± 5.5	80.7 ± 3.2	82.2 ± 3.1	0.646 ± 0.062
NeuralNetTorch_BAG_L1	81.3 ± 4.7	86.5 ± 4.6	84.0 ± 2.3	0.681 ± 0.046
NeuralNetTorch_BAG_L2	81.3 ± 3.8	89.5 ± 3.1	85.6 ± 3.0	0.711 ± 0.062
RandomForestEntr_BAG_L1	79.5 ± 4.8	89.4 ± 3.5	84.7 ± 3.2	0.694 ± 0.066
RandomForestEntr_BAG_L2	78.8 ± 4.6	89.8 ± 3.6	84.6 ± 3.0	0.693 ± 0.062
RandomForestGini_BAG_L1	79.6 ± 4.5	89.5 ± 4.1	84.8 ± 3.4	0.696 ± 0.071
RandomForestGini_BAG_L2	79.6 ± 5.0	89.6 ± 3.8	84.9 ± 3.3	0.698 ± 0.068
WeightedEnsemble_L2	80.7 ± 5.5	89.2 ± 3.3	85.2 ± 3.4	0.704 ± 0.069
WeightedEnsemble_L3	79.1 ± 4.4	91.4 ± 3.6	85.6 ± 3.0	0.714 ± 0.063
XGBoost_BAG_L1	79.3 ± 5.3	87.0 ± 2.8	83.4 ± 3.5	0.666 ± 0.071
XGBoost_BAG_L2	79.2 ± 3.9	90.0 ± 4.1	84.9 ± 2.9	0.698 ± 0.062

- Dataset from Jiang et al. 2019, Feng et al. 2021
- Reproductive toxicity
- 24 models with a variety of algorithms
- Consistent results with low standard deviations

Benchmarking results

- Models benchmarked against literature results
- Better accuracy than all results so far on this dataset used for benchmarking

Model	SE (%)	SP (%)	Accuracy (%)	MCC
Jiang et al. 2019	78.5	88.1	83.6	-
Feng et al. 2021	77.3	90.7	84.4	-
WeightedEnsemble_L3	79.1 ± 4.4	91.4 ± 3.6	85.6 ± 3.0	0.714 ± 0.063

Model	SE (%)	SP (%)	Accuracy (%)	MCC
CatBoost_BAG_L1	68.9 ± 2.6	73.5 ± 1.1	71.0 ± 1.0	0.423 ± 0.019
CatBoost_BAG_L2	69.1 ± 5.0	74.8 ± 4.2	71.7 ± 1.3	0.44 ± 0.025
ExtraTreesEntr_BAG_L1	73.0 ± 1.5	70.6 ± 1.9	71.8 ± 0.6	0.436 ± 0.012
ExtraTreesEntr_BAG_L2	71.8 ± 2.9	73.6 ± 2.9	72.6 ± 0.7	0.453 ± 0.014
ExtraTreesGini_BAG_L1	73.8 ± 2.4	69.3 ± 2.9	71.6 ± 1.0	0.431 ± 0.020
ExtraTreesGini_BAG_L2	71.8 ± 3.0	73.0 ± 3.2	72.4 ± 1.1	0.449 ± 0.024
LightGBMLarge_BAG_L1	71.9 ± 2.9	72.5 ± 1.7	72.2 ± 1.2	0.444 ± 0.023
LightGBMLarge_BAG_L2	69.2 ± 4.4	73.0 ± 3.8	71.0 ± 1.5	0.423 ± 0.028
LightGBMXT_BAG_L1	70.0 ± 2.9	72.7 ± 1.3	71.2 ± 1.2	0.426 ± 0.023
LightGBMXT_BAG_L2	70.3 ± 2.8	72.3 ± 2.6	71.2 ± 0.6	0.426 ± 0.013
LightGBM_BAG_L1	70.0 ± 2.9	72.7 ± 1.3	71.2 ± 1.2	0.426 ± 0.023
LightGBM_BAG_L2	70.2 ± 3.5	72.9 ± 3.2	71.5 ± 0.9	0.432 ± 0.017
NeuralNetFastAI_BAG_L1	69.2 ± 2.9	71.9 ± 3.1	70.5 ± 1.2	0.411 ± 0.024
NeuralNetFastAI_BAG_L2	69.9 ± 2.1	72.7 ± 2.8	71.2 ± 0.5	0.426 ± 0.013
NeuralNetTorch_BAG_L1	69.3 ± 2.2	71.5 ± 4.0	70.3 ± 1.1	0.408 ± 0.023
NeuralNetTorch_BAG_L2	71.4 ± 4.0	71.7 ± 3.1	71.5 ± 1.2	0.431 ± 0.024
RandomForestEntr_BAG_L1	72.9 ± 2.6	70.4 ± 2.8	71.6 ± 1.5	0.433 ± 0.030
RandomForestEntr_BAG_L2	70.6 ± 2.7	73.8 ± 3.6	72.1 ± 0.9	0.444 ± 0.018
RandomForestGini_BAG_L1	74.4 ± 2.5	69.7 ± 2.4	72.1 ± 1.4	0.442 ± 0.029
RandomForestGini_BAG_L2	70.9 ± 3.0	73.6 ± 3.2	72.1 ± 0.7	0.445 ± 0.015
WeightedEnsemble_L2	72.0 ± 2.1	71.6 ± 2.8	71.8 ± 0.6	0.436 ± 0.014
WeightedEnsemble_L3	70.2 ± 4.0	73.7 ± 2.6	71.8 ± 1.3	0.439 ± 0.026
XGBoost_BAG_L1	68.0 ± 3.1	74.5 ± 2.3	71.1 ± 0.8	0.426 ± 0.014
XGBoost_BAG_L2	69.7 ± 3.7	73.6 ± 3.6	71.5 ± 0.7	0.434 ± 0.015

Global models

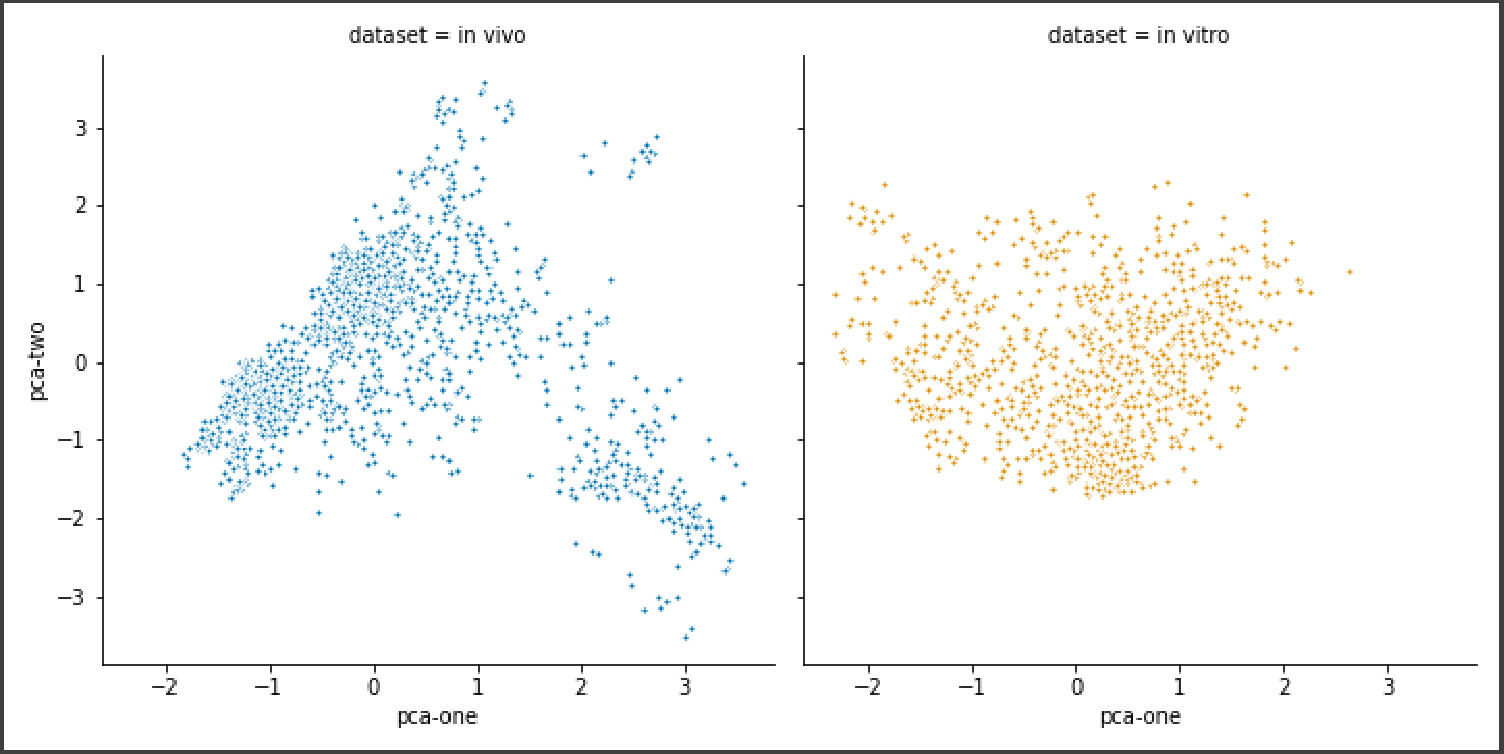
- Full database with 3255 compounds
- 24 models with a variety of algorithms
- Consistent results with low standard deviations
- Reasonable results given complexity of DART and lack of quality data

In vivo/in vitro results

- Models built for *in vivo* and *in vitro* datasets extracted from database
- Models built with *in vivo* data are much better

Test type	Model	SE (%)	SP (%)	Accuracy (%)	MCC
in vivo	RandomForestGini_BAG_L1	82.6 ± 1.1	80.1 ± 2.9	81.5 ± 1.6	0.625 ± 0.032
in vitro	RandomForestGini_BAG_L2	30.5 ± 6.0	89.2 ± 2.0	64.2 ± 3.8	0.246 ± 0.070

In vivo/in vitro results



- Principal component analysis (PCA) plots of *in vivo/in vitro* data
- Two different regions in feature space indicating chemicals are structurally different

Additional results

Comparison/Visualisation of other datasets extracted from database

- ML models and feature plots
- Developmental toxicity vs. reproductive toxicity
- By data source

Work in progress

Structural alerts (SAs)

- Associate structural fragment with mechanism of action for toxicity
- Constructed with Wedlake et al. 2020 KNIME workflow
- Comparison with SAs from DART scheme profile (Wu et al. 2013) in OECD QSAR Toolbox, DEREK SAs and other sources of SAs in the literature to see how many of these alerts are novel.
- SAs can be used for screening purposes or as part of next-generation risk assessments (NGRAs)

Conclusion

- We report the largest known database (ca. 3255 chemicals) for the general prediction of DART.
- We have used AutoML to quickly produce a best-performing model with 73% accuracy for predicting general DART.
- These results can be used for screening purposes or as part of next-generation risk assessments (NGRAs)

Acknowledgements

Goodman group

Unilever

Robinson College
University of Cambridge

UNIVERSITY OF
CAMBRIDGE

References

(1) Challa, A. P.; Beam, A. L.; Shen, M.; Peryea, T.; Lavieri, R. R.; Lippmann, E. S.; Aronoff, D. M. Machine learning on drug-specific data to predict small molecule teratogenicity. Reprod Toxicol 2020, 95, 148-158. DOI: 10.1016/j.reprotox.2020.05.004

(2) Ciallella, H. L.; Russo, D. P.; Sharma, S.; Li, Y.; Sloter, E.; Sweet, L.; Huang, H.; Zhu, H. Predicting Prenatal Developmental Toxicity Based On the Combination of Chemical Structures and Biological Data. Environ Sci Technol 2022, 56 (9), 5984-5998. DOI: 10.1021/acs.est.2c01040

(3) Di Filippo, J. I.; Bollini, M.; Cavasotto, C. N. A Machine Learning Model to Predict Drug Transfer Across the Human Placenta Barrier. Front Chem 2021, 9, 714678. DOI: 10.3389/fchem.2021.714678

(4) Evans, T. J.; Ganjam, V. K. Reproductive Anatomy and Physiology. In Reproductive and Developmental Toxicology, 2017; pp 7-37.

(5) Feng, H.; Zhang, L.; Li, S.; Liu, L.; Yang, T.; Yang, P.; Zhao, J.; Arkin, I. T.; Liu, H. Predicting the reproductive toxicity of chemicals using ensemble learning methods and molecular fingerprints. Toxicol Lett 2021, 340, 4-14. DOI: 10.1016/j.toxlet.2021.01.002

(6) Corvi, R.; Spielmann, H.; Hartung, T. Alternative Approaches for Carcinogenicity and Reproductive Toxicity. In The History of Alternative Test Methods in Toxicology, 2019; pp 209-217.

(7) Estevan, C.; Pamies, D.; Vilanova, E.; Sogorb, M. A. OECD Guidelines for In Vivo Testing of Reproductive Toxicity. In Reproductive and Developmental Toxicology, 2017; pp 163-178.

(8) Hartung, T.; Daneshian, M.; Hasiwa, N.; Leist, M. Validation and quality control of replacement alternatives – current status and future challenges. Toxicol Res (Camb) 2012, 1 (1), 8-22. DOI: 10.1039/c2tx20011b

(9) Vinardell, M. P. The use of non-animal alternatives in the safety evaluations of cosmetics ingredients by the Scientific Committee on Consumer Safety (SCCS). Regul Toxicol Pharmacol 2015, 71 (2), 198-204. DOI: 10.1016/j.yrtph.2014.12.018

(10) Sreedhar, D.; Manjula, N.; Pise, S. A.; Ligade, V. Ban of Cosmetic Testing on Animals: A Brief Overview. International Journal of Current Research and Review 2020, 12 (14), 113-116. DOI: 10.31782/ijcrr.2020.121424

References

(11) Gilmour, N.; Kimber, I.; Williams, J.; Maxwell, G. Skin sensitization: Uncertainties, challenges, and opportunities for improved risk assessment. Contact Dermatitis 2019, 80 (3), 195-200. DOI: 10.1111/cod.13167

(12) Baltazar, M. T.; Cable, S.; Carmichael, P. L.; Cubberley, R.; Cull, T.; Delagrange, M.; Dent, M. P.; Hatherell, S.; Houghton, J.; Kukic, P.; et al. A Next-Generation Risk Assessment Case Study for Coumarin in Cosmetic Products. Toxicol Sci 2020, 176 (1), 236-252. DOI: 10.1093/toxsci/kfaa048

(13) Kim, K. B.; Kwack, S. J.; Lee, J. Y.; Kacew, S.; Lee, B. M. Current opinion on risk assessment of cosmetics. J Toxicol Environ Health B Crit Rev 2021, 24 (4), 137-161. DOI: 10.1080/10937404.2021.1907264

(14) Wu, S.; Fisher, J.; Naciff, J.; Laufersweiler, M.; Lester, C.; Daston, G.; Blackburn, K. Framework for identifying chemicals with structural features associated with the potential to act as developmental or reproductive toxicants. Chem Res Toxicol 2013, 26 (12), 1840-1861. DOI: 10.1021/tx400226u

(15) Erickson, N. M., J.; Shirkov, A.; Zhang, H.; Larroy, P.; Li, M.; Smola, A. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. 2020. DOI: 10.48550/arxiv.2003.06505

(16) Jiang, C.; Yang, H.; Di, P.; Li, W.; Tang, Y.; Liu, G. In silico prediction of chemical reproductive toxicity using machine learning. Journal of Applied Toxicology 2019, 39 (6), 844-854. DOI: 10.1002/jat.3772

(17) Zurlinden, T. J.; Saili, K. S.; Rush, N.; Kothiya, P.; Judson, R. S.; Houck, K. A.; Hunter, E. S.; Baker, N. C.; Palmer, J. A.; Thomas, R. S.; et al. Profiling the ToxCast Library With a Pluripotent Human (H9) Stem Cell Line-Based Biomarker Assay for Developmental Toxicity. Toxicol Sci 2020, 174 (2), 189-209. DOI: 10.1093/toxsci/kfaa014

(18) Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. Journal of machine learning research 2008, 9 (11), 2579-2605.

(19) Hewitt, M.; Ellison, C. M.; Enoch, S. J.; Madden, J. C.; Cronin, M. T. Integrating (Q)SAR models, expert systems and read-across approaches for the prediction of developmental toxicity. Reprod Toxicol 2010, 30 (1), 147-160. DOI: 10.1016/j.reprotox.2009.12.003

(20) Baltazar, M.T., Cable, S., Carmichael, P.L., Cubberley, R., Cull, T., Delagrange, M., Dent, M.P., Hatherell, S., Houghton, J., Kukic, P. and Li, H., 2020. A next-generation risk assessment case study for coumarin in cosmetic products. Toxicological Sciences, 176 (1), 236-252.

(21) Wedlake, A. J.; Folia, M.; Piechota, S.; Allen, T. E. H.; Goodman, J. M.; Gutsell, S.; Russell, P. J. Structural Alerts and Random Forest Models in a Consensus Approach for Receptor Binding Molecular Initiating Events. Chem Res Toxicol 2020, 33 (2), 388-401. DOI: 10.1021/acs.chemrestox.9b00325