Leveraging High-Throughput Screening Data and Conditional Generative Adversarial Networks to Advance Predictive Toxicology

Adrian J. Green[1], Martin J. Mohlenkamp[2], Jhuma Das[3], Meenal Chaudhari[4], Lisa Truong[5], Robyn L. Tanguay[5], David M. Reif[1]

[1]Department of Biological Sciences, and the Bioinformatics Research Center, NC State University, Raleigh, NC; [2]Department of Mathematics, Ohio University, Athens, OH; [3]Marsico Lung Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC; [4]Department of Computational Science and Engineering, North Carolina A&T State University, Greensboro, NC; [5]Department of Environmental and Molecular Toxicology, Oregon State University, Corvallis, OR

¶ Corresponding Author:

David M. Reif, PhD

Ricks Hall 344, Box 7566,

1 Lampe Drive, North Carolina State University,

Raleigh, NC 27695, USA.

Telephone: 919-513-3812.

E-mail Address: dmreif@ncsu.edu

Acknowledgements

This research was supported by the National Institute of Environmental and Health Sciences at the National Institutes of Health and the Statistical and Applied Mathematical Sciences Institute.

SUMMARY

A conditional Generative Adversarial Network (cGAN) can leverage a large chemical set of experimental toxicity data plus chemical structure information to predict the toxicity of untested compounds.

ABSTRACT

There are currently 85,000 chemicals registered with the EPA under the Toxic Substances Control Act, but only a small fraction (~5%) have any measured toxicological data. To address this data gap, high-throughput screening (HTS) methods are vital. As part of one such HTS effort, embryonic zebrafish were used to examine a suite of morphological and mortality endpoints at six concentrations from over 1,000 unique chemicals found in the ToxCast library (phase 1 and 2). We hypothesized that by using a conditional Generative Adversarial Network (cGAN) and leveraging this large set of toxicity data, plus chemical structure information, we could efficiently predict toxic outcomes of untested chemicals. CAS numbers for each chemical were used to generate textual files containing three-dimensional structural information for each chemical in the Protein Data Bank (PDB) file format. Utilizing a novel method in this space, we converted the 3D structural information into a weighted set of points while retaining all information about the structure. The *in vivo* toxicity and chemical data were used to train two neural network generators. The first used regression to train a generator (Go-ZT) to produce toxicity data while the second utilized cGAN architecture to train a generator (GAN-ZT). Our results showed that both Go-ZT and GAN-ZT models produce similar results, but the cGAN achieved higher sensitivity (SE) value of 85.7% vs 71.4%. Conversely, Go-ZT attained a higher specificity (SP), positive predictive value (PPV), and Kappa results of 67.3%, 23.4%, and 0.21 compared to 24.5%, 14.0%, and 0.03 for the cGAN, respectively. By combining both Go-ZT and GAN-ZT, our consensus model improved the SP, PPV, and Kappa, to 75.5%, 25.0%, and 0.211, respectively, resulting in an area under the receiver operating characteristic (AUROC) of 0.663. Considering their potential use as efficient prescreening tools, these new models can provide *in vivo* toxicity predictions and insight into as-yet untested areas of chemical space to prioritize compounds for HT testing.

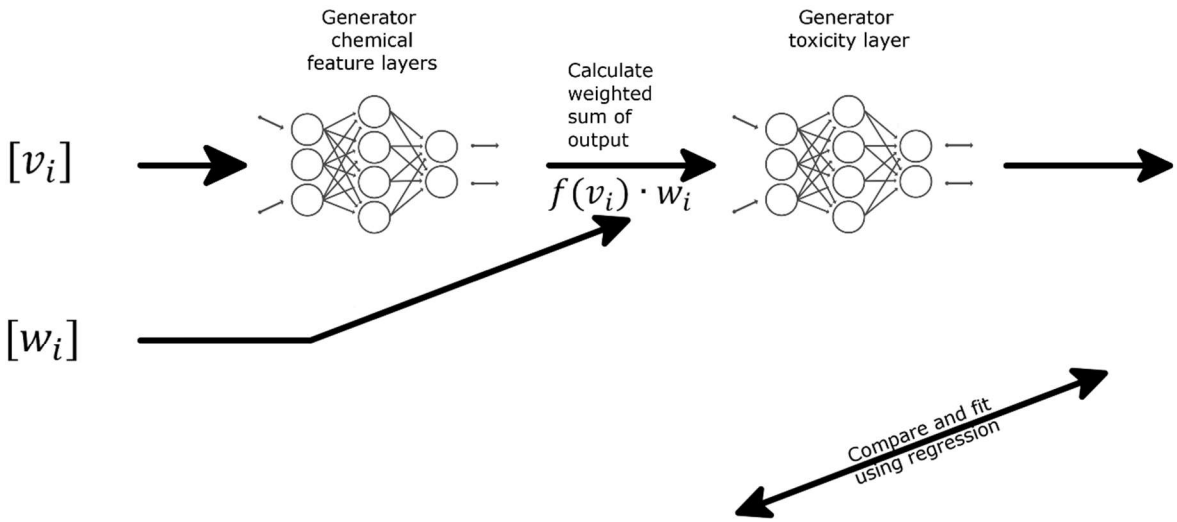
INTRODUCTION

Currently, there are 85,000 chemicals registered with the EPA, as part of the Toxic Substances Control Act[1], that are manufactured, processed, or imported into the United States; however, only 4,400 have rigorous toxicological data, leaving over 80,000 chemicals untested [2,3]. Due to the high cost, and ethical concerns over the use of low-throughput mammalian models associated with traditional *in vitro* and *in vivo* assays, there has been increasing demand to reduce the number of animals used in toxicity testing paradigms by switching to *in silico* methods [4]. To directly address this chemical data gap and help prioritize chemicals for testing, the ToxCast program was developed[5,6]. The original ToxCast program included approximately 700 biochemical and cell-based assays, which was efficient but lacking in systemic biological complexity[2]. Therefore, as part of an effort to expand the toxicology database, a multidimensional, high-throughput screening (HTS) assay was devised to examine all ToxCast phase 1 and 2 chemicals (over 1,000 unique chemicals) for developmental and neuro-toxicity in the embryonic zebrafish [7].

Additionally, computational approaches to bridge the data gap above have been employed, with Quantitative Structure-Activity Relationship (QSAR) and Read-Across being the most commonly used methodologies [8–13]. Both methods rely on the grouping of chemicals together using functional features, e.g. number of carbons, and have employed statistical or machine learning approaches. Although these methods have been useful in identifying priority compounds for further testing, how these chemicals are grouped together might add bias and recent machine learning advances have not been thoroughly explored [14]. Therefore, we propose using a supervised generative adversarial machine learning architecture that leverages the zebrafish HTS assay data along with chemical structure information to predict the toxic outcomes of untested chemicals. This approach will expand QSAR methodologies by providing a computational toxicology tool capable of prescreening the untested chemical space for active developmental toxicants.

MATERIALS AND METHODS

In this section, we describe a new conditional generative adversarial network utilizing a novel weighted sum over views notation of compounds to predict active developmental toxicants. An overview of our approach is shown in Fig. 1 and 2. First, we used experimental data collected on a large, diverse compound set to assess the toxic effects of these chemicals following developmental exposure (Fig 3a). Next, we recast the chemical data in a structural representation that maintained connectivity and positional information of each atom in the molecule but in a format easily read as input into a neural network. Next, we trained two types of generators to produce toxicity data using the recast chemical structural representation. The first used a cGAN architecture while the second utilized a deep neural network (DNN) with regression for training. This was done to produce generators capable of predicting developmental zebrafish toxicity data dependent on chemical structure alone. Regression training was leveraged to maximize model fit, minimize training time, and utilized a simpler toxicity data representation. cGAN training minimized the effects of outliers and increased network adaptability to chemical structure. These generators were trained using phase 1 and 2 ToxCast chemical data (n = 1003) split 80:20 into training and validation sets (Fig 3a). Finally, we evaluated the trained networks on an independent test set containing chemicals (n = 57) of greater diversity in terms of both size and atomic constituents (Fig 3b).



Biologocal State	Abbreviation	0 µM	0.0064 µM	0.064 µM	0.64 µM	6.4 µM	64 µM
Mortality	MORT	6	6	2	3	6	3
Yolk Sac Edema	YSE_	1	4	2	7	3	2
Body Axis	AXIS	1	2	1	16	24	28
Eye Defect	EYE_	1	4	1	4	1	2
Snout	SNOU	1	6	10	19	24	28
Jaw	JAW_	1	7	10	19	24	28
Otic Vesicle	OTIC	1	4	1	4	1	13
Pericardial Edema	PE__	1	4	2	6	2	2
Brain	BRAI	1	4	1	15	24	28
Somite	SOMI	0	0	0	0	0	0
Pectoral Fin	PFIN	1	4	1	16	24	28
Caudal Fin	CFIN	1	4	2	4	2	3
Pigment	PIG_	1	19	16	27	24	28
Circulation	CIRC	0	0	0	0	0	0
Truncated Body	TRUN	1	4	1	5	2	3
Swim Bladder	SWIM	0	0	1	0	0	0
Notochord & Bent Tail	NC__	0	0	0	0	0	0
Touch Response	TR__	1	1	1	3	1	3

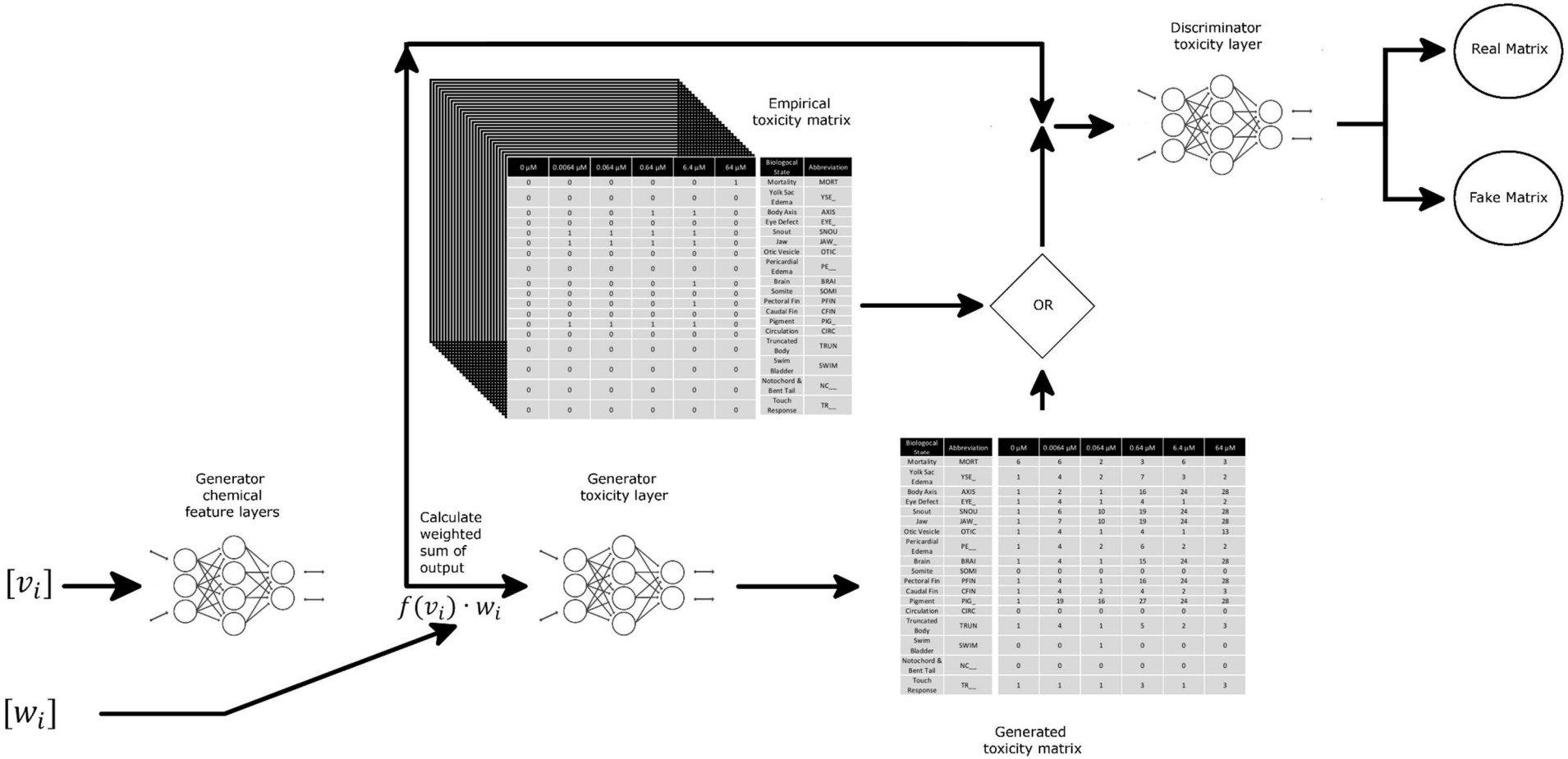
Generated toxicity matrix

Empirical toxicity matrix

Biologocal State	Abbreviation	0 µM	0.0064 µM	0.064 µM	0.64 µM	6.4 µM	64 µM
Mortality	MORT	6	6	2	3	6	3
Yolk Sac Edema	YSE_	1	4	2	7	3	2
Body Axis	AXIS	1	2	1	16	24	28
Eye Defect	EYE_	1	4	1	4	1	2
Snout	SNOU	1	6	10	19	24	28
Jaw	JAW_	1	7	10	19	24	28
Otic Vesicle	OTIC	1	4	1	4	1	13
Pericardial Edema	PE__	1	4	2	6	2	2
Brain	BRAI	1	4	1	15	24	28
Somite	SOMI	0	0	0	0	0	0
Pectoral Fin	PFIN	1	4	1	16	24	28
Caudal Fin	CFIN	1	4	2	4	2	3
Pigment	PIG_	1	19	16	27	24	28
Circulation	CIRC	0	0	0	0	0	0
Truncated Body	TRUN	1	4	1	5	2	3
Swim Bladder	SWIM	0	0	1	0	0	0
Notochord & Bent Tail	NC__	0	0	0	0	0	0
Touch Response	TR__	1	1	1	3	1	3

Figure 1. Regression Generator Diagram.

Schematic representation of Go-ZT architecture showing chemical structural input represented as weights (w_i) and views (v_i) matrices passed through two fully connected neural networks to produce a predicted toxicity matrix.



0 µM	0.0064 µM	0.064 µM	0.64 µM	6.4 µM	64 µM	Biologocal State	Abbreviation
0	0	0	0	0	1	Mortality	MORT
0	0	0	0	0	0	Yolk Sac Edema	YSE_
0	0	0	1	1	0	Body Axis	AXIS
0	0	0	0	0	0	Eye Defect	EYE_
0	1	1	1	1	0	Snout	SNOU
0	1	1	1	1	0	Jaw	JAW_
0	0	0	0	0	0	Otic Vesicle	OTIC
0	0	0	0	0	0	Pericardial Edema	PE__
0	0	0	0	1	0	Brain	BRAI
0	0	0	0	0	0	Somite	SOMI
0	0	0	0	1	0	Pectoral Fin	PFIN
0	0	0	0	0	0	Caudal Fin	CFIN
0	1	1	1	1	0	Pigment	PIG_
0	0	0	0	0	0	Circulation	CIRC
0	0	0	0	0	0	Truncated Body	TRUN
0	0	0	0	0	0	Swim Bladder	SWIM
0	0	0	0	0	0	Notochord & Bent Tail	NC__
0	0	0	0	0	0	Touch Response	TR__

Biologocal State	Abbreviation	0 µM	0.0064 µM	0.064 µM	0.64 µM	6.4 µM	64 µM
Mortality	MORT	6	6	2	3	6	3
Yolk Sac Edema	YSE_	1	4	2	7	3	2
Body Axis	AXIS	1	2	1	16	24	28
Eye Defect	EYE_	1	4	1	4	1	2
Snout	SNOU	1	6	10	19	24	28
Jaw	JAW_	1	7	10	19	24	28
Otic Vesicle	OTIC	1	4	1	4	1	13
Pericardial Edema	PE__	1	4	2	6	2	2
Brain	BRAI	1	4	1	15	24	28
Somite	SOMI	0	0	0	0	0	0
Pectoral Fin	PFIN	1	4	1	16	24	28
Caudal Fin	CFIN	1	4	2	4	2	3
Pigment	PIG_	1	19	16	27	24	28
Circulation	CIRC	0	0	0	0	0	0
Truncated Body	TRUN	1	4	1	5	2	3
Swim Bladder	SWIM	0	0	1	0	0	0
Notochord & Bent Tail	NC__	0	0	0	0	0	0
Touch Response	TR__	1	1	1	3	1	3

Figure 2. Conditional GAN Diagram.

Schematic representation of GAN-ZT architecture showing chemical structural input represented as weights (w_i) and views (v_i) matrices passed through two fully connected neural networks to produce a predicted toxicity matrix. Chemical features along with predicted or empirical toxicity matrices are then passed to a discriminator comprising a fully-connected neural network.

Empirical (Experimental) Data

The empirical data used to develop a generator of zebrafish toxicity were gathered as described in Truong et al. and Noyes et al.[7,15] Figure 4 shows the experimental design and input data structure. The data included 1003 unique ToxCast chemicals tested at six concentrations for each chemical (0 µM, 0.0064 µM,0.064 µM, 0.64 µM, 6.4 µM and 64 µM). To minimize effects of response variability at lower concentrations, only the highest concentration was chosen for network training. There were 32 replicates (an individual embryo in singular wells of a 96-well plate) at each concentration for each chemical. At 120 hours post-fertilization (hpf), 18 distinct developmental endpoints were evaluated. The data were recorded as binary incidences and split into training (n = 802) and validation (n = 201) chemical sets (Fig 3a). The data were either summed for each individual and normalized by the number of larvae per concentration or kept as individual 1 by 18 matrices and used to train, and validate our regression and GAN networks, respectively. In a similar manner, toxicity matrices were created for an independent test set of 57 chemicals that were collected in new experiments after collection of the original ToxCast data. This new test chemical set was more diverse in terms of atomic species and physical chemical properties (Fig 3b)[16]. Due to chemical vectorization constraints we defined a reasonable domain of applicability to exclude, Perfluorinated chemicals with carbon chains longer than nine, Chloroperfluoro chemicals, and chemicals with a betaine functional group. Figure 3 shows the division of these data into train/validate/test subsets and PCA comparisons of the physicochemical properties using the Environmental Protection Agency's (EPA) Integrated Chemical Environment (ICE) chemical characterization tool[16]. The PCA analysis includes the following physical chemical properties Molecular Weight, Boiling Point, Henry's Law, Constant Melting Point, Negative Log of Acid Dissociation Constant, Octanol-Air Partition Coefficient, Octanol-Water Distribution Coefficient, Octanol-Water Partition Coefficient, Vapor Pressure, Water Solubility [16].

Representing Chemical Compounds

Utilizing their CAS numbers chemical structural information was retrieved from the EPA's Chemistry Dashboard and converted from SDF to PDB format using Open Babel[17,18]. The PDB format was chosen as it is easily accessible and contains 3D structural information for all atoms in a molecule. Multiple methods have been used to encode this structural information for

utilization in deep learning, including, chemical properties, molecular fingerprints, SMILES, and graph vectorization, as well as 2D images of a chemical [19–21]. In this analysis, we utilized an algorithm developed to map and vectorize structure that was originally developed for use in material sciences [22]. The PDB file for each chemical was vectorized as described by d'Avazac et. al [22] and illustrated in Fig. s1. This method is simple and universal with few parameters and was adapted as follows; First, each chemical was vectorized (x, y, and z coordinates) using either all atoms or only carbon atoms, when available, in the PDB list. The relative position of each additional atom in the PDB list was added to the view until all atoms were added or a user defined limit was reached. Lastly, an

atom's position on the periodic table was used as a unique identifier (group and period) producing five chemical features per atom.

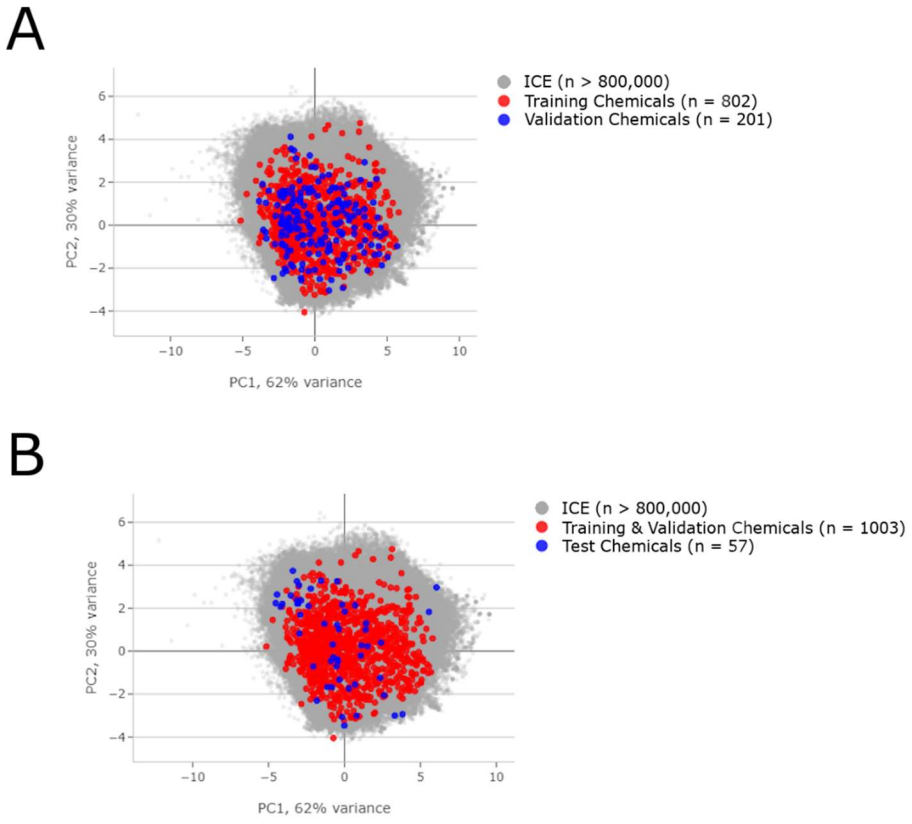


Figure 3. Data Subdivision.

Principal component analysis displayed against the background of over 800,000 chemicals in the EPA Integrated Chemical Environment (ICE) database. (A) Compares physical chemical properties between the training and validation sets. (B) Compares physical chemical properties between the training and validation set, and the test set.

cGAN and Regression Generator

Two network architectures were developed and tested to train a generator that was capable of using 3D chemical information, in the form of vectorize views (Fig s1), and generate a toxicity matrix (Fig 1 and 2). The following two different models were trained on a Dell R740 containing two Intel Xeon processors with 18 cores per processor, 512 GB RAM, and Tesla-V100-PCIE (31.7 GB). The first was a simpler deep neural network trained to produce a toxicity matrix using regression (Fig 1). We used multiple layers consisting of a deep neural network base layer to extract salient features from the views matrix for each chemical structure (generator chemical feature layer). A second layer calculated the weighted sum over features ($f(v_i) \cdot w_i$) and a final deep neural network (generator toxicity layer) generates toxicity values. The regression generator (Go-ZT) was trained over the course of 200 epochs (35 seconds/run) and the generated toxicity matrices were compared to empirical toxicity matrices and the network that generated toxicity matrices with the highest Cohen's Kappa statistic was used for further evaluation.

The second network developed (GAN-ZT) used a much more complex cGAN architecture to generate a toxicity matrix (Fig. 2). Similar to the Go-ZT above, we used multiple layers consisting of a deep neural network base layer to extract salient features from the views matrix for each chemical structure (generator chemical feature layer) and a second layer to calculate the weighted sum over these features ($f(v_i) \cdot w_i$). The resulting weighted sum of views($f(v_i) \cdot w_i$) for each chemical was used as input to a final deep neural network (generator toxicity layer) to produce a generated toxicity matrix. The discriminator took the generators resulting weighted sum of views($f(v_i) \cdot w_i$) for each chemical along with its corresponding empirical or generated toxicity matrices to determine whether the toxicity matrix was real or fake. This information is then backpropagated to train the generator. GAN-ZT was trained over the course of 500 epochs (1.5 hours/run) and the network that generated toxicity matrices with the highest Cohens Kappa statistic was used for further evaluation.

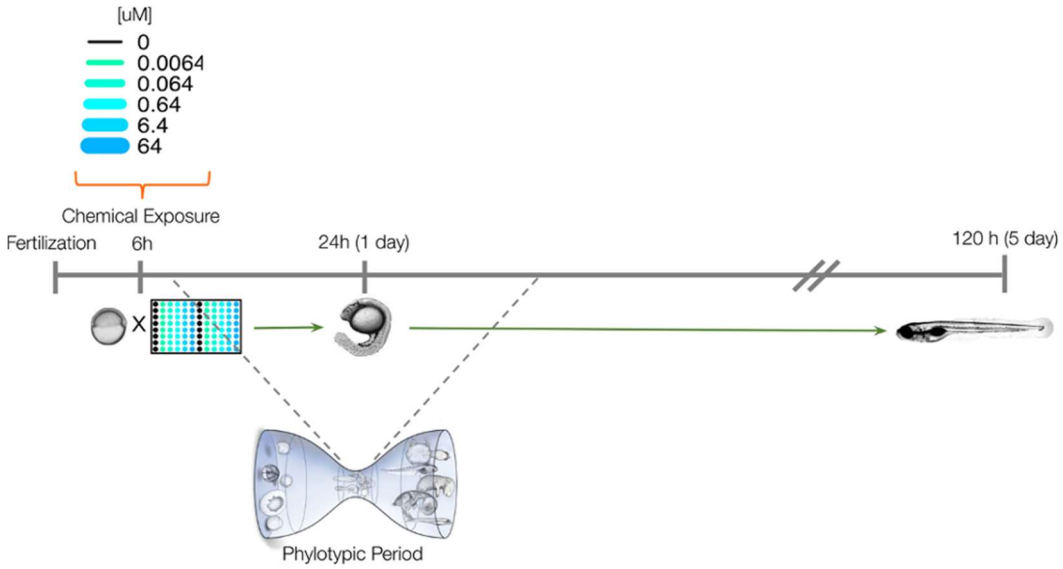


Figure 4. Experimental Design.

Schematic representation of the experimental approach for screening developmental and neurotoxicity of chemicals in larval zebrafish.

Network Performance and Evaluation

Previous work analyzing this data has shown that a summary value, aggregate entropy (AggE) can be calculated from the 18 morphological endpoints [23]. The intent of this summary value is to capture a meaningful measure of toxicity, while avoiding overinflation by summing highly correlated endpoints. Using the threshold value (9.35) identified by Zhang et. al for AggE in these data, compounds may be classified as active or inactive[23]. It should be noted that this threshold value influences the toxicity hit rate of a chemical and is concentration dependent. Therefore, it would need to be changed when investigating other nominal concentration. Following training, the resulting generators were used to output toxicity matrices for the 201 chemicals in the validation dataset. AggE values were calculated using both the empirical and generated toxicity matrices and compounds were classified as active or inactive using the threshold value identified above. Active vs inactive classification accuracy was evaluated using a confusion matrix, Cohen's Kappa statistic, and area under the receiver operating characteristic (AUROC). The primary metrics we used from the confusion matrix included sensitivity (SE), specificity (SP), and positive predictive value (PPV) as these give us the true positive rate, true negative rate, and the proportion of true positives amongst all positive calls[24–26]. While Kappa and AUROC measure model accuracy, while compensating for simple chance [27]. The network with the highest Kappa statistic and positive predictive value (PPV) was used for evaluation of the test dataset.

RESULTS

Capacity Ceiling and Network training

The first step in training both Go-ZT and GAN-ZT included recasting the chemical structures as views and extracting relevant features from this chemical data. Therefore, at the very least our networks should be capable of extracting physical properties from the chemical data provided and gives a floor to the potential performance of our network. As shown in Fig. s2, our GAN-ZT architecture generates chemical diameters with a Mean Absolute Percent Error (MAPE) of 12.4% and 13.0% for the training and validation datasets, respectively. While Go-ZT architecture generates chemical diameters with a MAPE of 10.3% and 13.8% for the training and validation datasets, respectively. Therefore, the best performance we could expect from either GAN-ZT or Go-ZT is approximately 13.0%.

Following training using chemical and toxicity data the best networks utilized Swish activation, batch normalization between layers, mean squared error for the loss function, and Adam as the optimizer[28–30]. GAN-ZT and Go-ZT generated toxicity values, using the chemical structures found in our training and validation datasets, with a validation MAPE of 49.1% and 51.6%, respectively (Fig. s3).

Predicting Toxicity

Empirical and generated toxicity matrices for each chemical from the validation or test datasets (Fig. 3b) were used to calculate AggE values and determine activity classification. The empirical validation dataset contains 22 chemicals that meet the AggE threshold to be classified as active compounds (10.9%). As shown in Table 1, changing the number of chemical views from 43 (preferring views starting with Carbon) to 126 (maximum), decreasing the maximum number of atoms per view from 118 (maximum number of atoms in the training set) to 20 by truncation, and decreasing number of chemical features from 590 (maximum) to 120 in training altered the SE, Kappa, and AUROC results of both the Go-ZT and GAN-ZT network architectures. GAN-ZT_v2 and Go-ZT_v3 were chosen for final evaluation using the test dataset as they had the best Kappa and AUROC values.

The Go-ZT_v3 and GAN-ZT_v2 networks predicted 5 and 15 of the 22 active chemicals in the validation dataset, respectively. As shown in Table 1 this resulted in a SE, SP, and PPV for Go-ZT of 22.7%, 91.6%, and 25.0% respectively. GAN-ZT on the other hand produced SE, SP, and PPV values of 68.2%, 60.3%, and 17.4%, respectively. Evaluation of the chemical domain space

using Go-ZT and GAN-ZT showed that the chemicals excluded due to long chain lengths, or betaine or Chloroperfluoro functional groups showed that these chemical properties fall outside of the domain space of our models.

Go-ZT and GAN-ZT were then used to predict the toxicity of an independent and more diverse chemical set containing seven active compounds (12.5%). The results (Table 2) show that Go-ZT performed best with increases in both Kappa and AUROC values while GAN-ZT saw declines in their Kappa and AUROC values. Overall Go-ZT had higher ACC, SP, and PPV while GAN-ZT had a better SE (fig. 5). By combining the predictive results of these two machine learning architectures we were able to leverage the sensitivity of the cGAN and the specificity of the regression model (fig. 6). As a result of the consensus between the models we were able to capture four of the seven active chemicals in the test set while eliminating 29 false positives which translates to a PPV of 25.0%.

Table 1. Prediction Performances Validation Chemicals.

Network	Number of Conc.	Atoms per view	Chemical Views	SE	SP	PPV	Kappa	AUROC
GAN-ZT_v1	1	118	43	40.9	73.2	15.8	0.083	0.5705
Go-ZT _v1	1	118	43	62.5	43.5	13.0	0.023	0.5300
GAN-ZT_v2	1	118	126	68.2	60.3	17.4	0.125	0.6426
Go-ZT _v2	1	118	126	37.5	81.4	21.4	0.142	0.5943
GAN-ZT_v3	1	24	126	45.5	72.6	16.9	0.104	0.5904
Go-ZT_v3	1	24	126	22.7	91.6	25.0	0.149	0.5717

Table 2. Prediction Performances Test Chemicals.

Network	Number of Conc.	Atoms per view	Chemical Views	SE	SP	PPV	Kappa	AUROC
GAN-ZT_v2	1	118	126	85.7	24.5	14.0	0.032	0.5510
Go-ZT_v3	1	24	126	71.4	67.3	23.8	0.209	0.6939

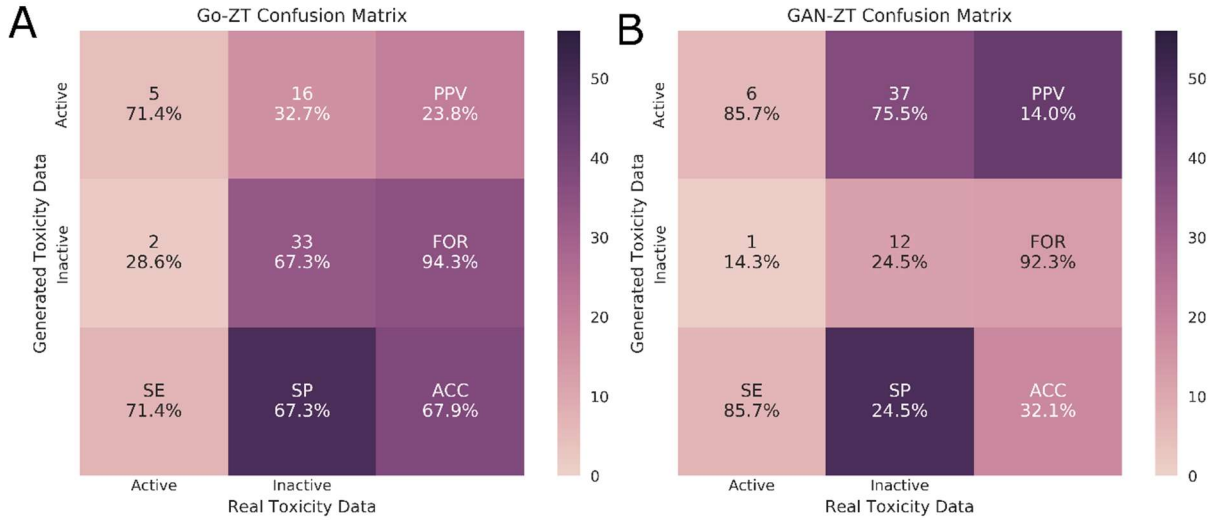


Figure 5. Test Dataset Confusion Matrices.
Evaluation of the classification of chemicals in the test data set as either active or inactive using real versus generated toxicity matrices by Go-ZT_v3 (A) or GAN-ZT_v2 (B). PPV – positive predictive value, FOR – false omission rate, SE – sensitivity, SP – specificity, ACC – overall accuracy.

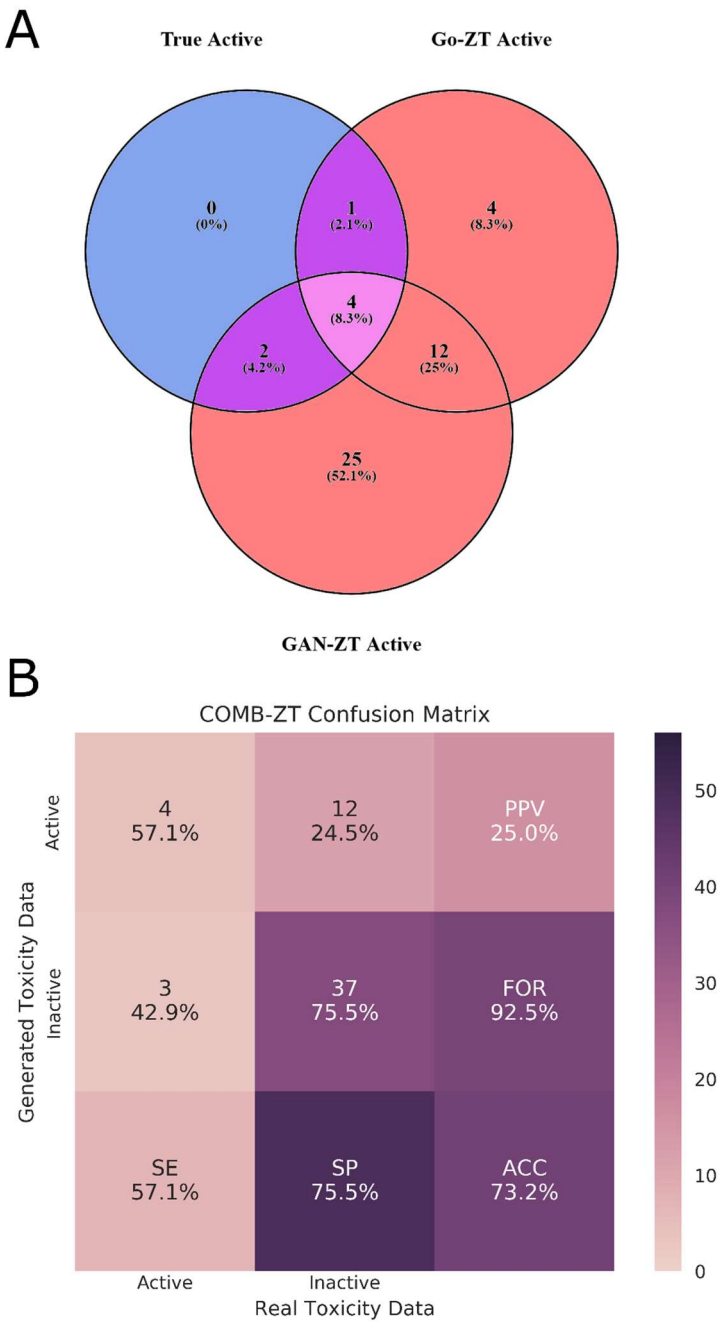


Figure 6. Model Consensus on Chemical Activity.
Venn diagram showing the overlap between true active chemicals and chemicals predicted to be active by either Go-ZT_v3 or GAN-ZT_v2 (A). A confusion matrix showing the performance measures of the combined Go-ZT_v3 and GAN-ZT_v2 models (B) using the test dataset.

DISCUSSION

GAN-ZT and Go-ZT architecture with chemical structure vectorization using views predicts empirical toxicity results with fair Kappa values of 0.125 and 0.149, respectively. Evaluation of an independent and more diverse chemical test set showed similar Kappa, AUROC, and PPV values for Go-ZT while GAN-ZT Kappa, AUROC, and PPV performance declined. Go-ZT_v3 predicted active chemicals with an SE of 71.4%, SP of 67.3%, and a PPV of 23.8% while GAN -ZT_v2 predicted active chemicals with an SE of 85.7%, SP of 24.5%, and a PPV of 14.0%. When we examined the overlap in predicted active chemicals between Go-ZT and GAN-ZT we see that a consensus model improved the SP, PPV, and Kappa, to 75.5%, 25.0%, and 0.211, respectively but decreased the AUROC slightly to 0.663.

These results show that both supervised generative adversarial and regression-based machine learning architectures are capable of predicting toxic developmental activity with fair to moderate efficacy, especially when we consider the diversity of the test set. Further, leveraging the strengths of both architectures the intersection between the models was able to accurately predict the toxicity of chemicals not part of the initial ToxCast screen.

Other machine learning methods such as Deep Neural Networks (DNN), Support Vector Machines (SVM), *k*-nearest neighbor (*k*-NN), gradient-boosted decision trees, and Bayesian Classifiers have been applied to cheminformatics problems to predict biologically active chemicals with AUROC values ranging from 0.7 – 0.83[31]. These studies utilized ECFP4 fingerprints of chemicals from the ChEMBL database, and single biological activity prediction in drug discovery but were not focused on *in vivo* toxicity. More recent studies have used data from the Tox21 database as part of the NIH 2014 Tox21 Data Challenge with multitask DNNs outperforming other machine learning methods with AUROC values ranging from 0.69 – 0.92 in 12 different biochemical assays[20]. Further, Mansouri and Judson successfully build a QSARs model for G-protein coupled receptor assays using partial least square discriminant analysis that resulted in a balanced accuracy of 96%[32]. Our combined model produced a similar AUROC (0.663) and a lower balanced accuracy value (66%). To the best of our knowledge, this is the first study to develop a DNN model without explicit use of molecular descriptors to predict *in vivo* toxicity in a large chemical set.

While our model is a potentially useful tool for prioritizing chemicals for screening tests, it does have its limitations including high computational costs for training and chemical domain space limits. A better generator could be developed with additional resources. In addition, our networks are dependent on accurate 3D chemical structural information to produce reliable results and at this time are not designed to evaluate mixtures.

Overall, our results show that a DNN utilizing 3D chemical structural information is a useful prescreening tool for predicting the toxic outcomes of the approximately 80,000 untested chemicals registered with the EPA. If we consider that between the training, validation, and test set there are 1,060 chemicals and only 145 are active (13.7%) then there are possibly 10,900 untested active chemicals registered. If the PPV holds at 25% this would result in a list of ~45,000 chemicals to screen. While still a very large number it would reduce the experimental space by over half. Further, these compounds may then be ranked by AggE and the highest chemicals identified should then be prioritized for assessment using the zebrafish HTS assay for developmental toxicity as these assays are considerably faster and cheaper than tradition chemical screens in mammalian systems

Looking to the future, increasing computational resources and chemical structural data, alternative network architectures, and inclusion of ToxCast assay results and zebrafish behavioral endpoints in conditional training could improve the predictive value of DNN in *in vivo* toxicity testing. Additional work needs to be done to assess the utility of GAN architecture as a tool to evaluate structure activity relationships (SAR) in toxicology and finally DNNs need to be adapted to evaluate mixtures if a sufficiently large dataset is available.

REFERENCES

1. US EPA, O. About the TSCA Chemical Substance Inventory. *US EPA* https://www.epa.gov/tsca-inventory/about-tsca-chemical-substance-inventory (2015).
2. US EPA, O. ToxCast Chemicals. *US EPA* https://www.epa.gov/chemical-research/toxcast-chemicals (2017).
3. Richard, A. M. *et al.* ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chem. Res. Toxicol.* **29**, 1225–1251 (2016).
4. Krewski, D. *et al.* TOXICITY TESTING IN THE 21ST CENTURY: A VISION AND A STRATEGY. *J Toxicol Environ Health B Crit Rev* **13**, 51–138 (2010).
5. Judson, R. S. *et al.* In vitro screening of environmental chemicals for targeted testing prioritization: the ToxCast project. *Environ. Health Perspect.* **118**, 485–492 (2010).
6. Dix, D. J. *et al.* The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol. Sci.* **95**, 5–12 (2007).
7. Truong, L. *et al.* Multidimensional In Vivo Hazard Assessment Using Zebrafish. *Toxicol Sci* **137**, 212–233 (2014).
8. Matsuzaka, Y. & Uesawa, Y. DeepSnap-Deep Learning Approach Predicts Progesterone Receptor Antagonist Activity With High Performance. *Front. Bioeng. Biotechnol.* **7**, (2020).
9. Idakwo, G. *et al.* Deep Learning-Based Structure-Activity Relationship Modeling for Multi-Category Toxicity Classification: A Case Study of 10K Tox21 Chemicals With High-Throughput Cell-Based Androgen Receptor Bioassay Data. *Front. Physiol.* **10**, (2019).
10. Pawar, G., Madden, J. C., Ebbrell, D., Firman, J. W. & Cronin, M. T. D. In Silico Toxicology Data Resources to Support Read-Across and (Q)SAR. *Front. Pharmacol.* **10**, (2019).

11. Yoo, J. W. *et al.* Development of improved QSAR models for predicting the outcome of the in vivo micronucleus genetic toxicity assay. *Regulatory Toxicology and Pharmacology* **113**, 104620 (2020).
12. Ghorbanzadeh, M., Zhang, J. & Andersson, P. L. Binary classification model to predict developmental toxicity of industrial chemicals in zebrafish. *Journal of Chemometrics* **30**, 298–307 (2016).
13. Zhang, H. *et al.* Development of novel in silico model for developmental toxicity assessment by using naïve Bayes classifier method. *Reproductive Toxicology* **71**, 8–15 (2017).
14. Non-test Methods (Q)SAR and Read-across. *AltTox.org* http://alttox.org/mapp/emerging-technologies/non-test-approaches-qsars-read-across/ (2014).
15. Noyes, P. D., Haggard, D. E., Gonnerman, G. D. & Tanguay, R. L. Advanced Morphological — Behavioral Test Platform Reveals Neurodevelopmental Defects in Embryonic Zebrafish Exposed to Comprehensive Suite of Halogenated and Organophosphate Flame Retardants. *Toxicol Sci* **145**, 177–195 (2015).
16. National Toxicology Program. ICE Tools. https://ice.ntp.niehs.nih.gov/Tools (2020).
17. O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **3**, 33 (2011).
18. Williams, A. J. *et al.* The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *Journal of Cheminformatics* **9**, 61 (2017).
19. Hirohara, M., Saito, Y., Koda, Y., Sato, K. & Sakakibara, Y. Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. *BMC Bioinformatics* **19**, 526 (2018).

20. Mayr, A., Klambauer, G., Unterthiner, T. & Hochreiter, S. DeepTox: Toxicity Prediction using Deep Learning. *Front. Environ. Sci.* **3**, (2016).
21. Guimaraes, G. L., Sanchez-Lengeling, B., Outeiral, C., Farias, P. L. C. & Aspuru-Guzik, A. Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models. *arXiv:1705.10843 [cs, stat]* (2018).
22. d'Avezac, M., Botts, R., Mohlenkamp, M. J. & Zunger, A. Learning to Predict Physical Properties using Sums of Separable Functions. *SIAM J. Sci. Comput.* **33**, 3381–3401 (2011).
23. Zhang, G., Marvel, S., Truong, L., Tanguay, R. L. & Reif, D. M. Aggregate entropy scoring for quantifying activity across endpoints with irregular correlation structure. *Reprod. Toxicol.* **62**, 92–99 (2016).
24. Pearson, K. On the theory of contingency and its relation to association and normal correlation. in *Drapers Company Research Memoirs* (Dulau and Co., 1904).
25. Townsend, J. T. Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics* **9**, 40–50 (1971).
26. Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G. & Thomas, R. Understanding and using sensitivity, specificity and predictive values. *Indian J Ophthalmol* **56**, 45–50 (2008).
27. Ben-David, A. About the relationship between ROC curves and Cohen's kappa. *Engineering Applications of Artificial Intelligence* **21**, 874–882 (2008).
28. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]* (2017).
29. Ramachandran, P., Zoph, B. & Le, Q. V. Searching for Activation Functions. *arXiv:1710.05941 [cs]* (2017).

30. Osl, M., Netzer, M., Dreiseitl, S. & Baumgartner, C. Applied Data Mining: From Biomarker Discovery to Decision Support Systems. in *Computational Medicine* (ed. Trajanoski, Z.) 173–184 (Springer Vienna, 2012). doi:10.1007/978-3-7091-0947-2_10.
31. Goh, G. B., Hodas, N. O. & Vishnu, A. Deep learning for computational chemistry. *Journal of Computational Chemistry* **38**, 1291–1307 (2017).
32. Mansouri, K. & Judson, R. S. In Silico Study of In Vitro GPCR Assays by QSAR Modeling. in *In Silico Methods for Predicting Drug Toxicity* (ed. Benfenati, E.) 361–381 (Springer, 2016). doi:10.1007/978-1-4939-3609-0_16.