Machine learning in predictive toxicology: investigating developmental and reproductive toxicity with transfer learning

by

Wang Wei How Marcus (mwhw3)

Robinson College

Supervisor: Professor Jonathan M Goodman

This thesis is submitted for the degree of Doctor of Philosophy in Chemistry

Yusuf Hamied Department of Chemistry

University of Cambridge

Submitted 07/2023

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted, or, is being concurrently submitted, for any degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Abstract

Machine learning in predictive toxicology: investigating developmental and reproductive toxicity with transfer learning

Wang Wei How Marcus

The toxicity of a compound has always been a major concern in risk assessments of new products or drugs. In the field of predictive toxicology, with animal testing being phased out in some sectors, there is an urgent need for alternative methods for determining toxicity. An *in silico* method such as machine learning is one such popular choice given that the results can be obtained quickly with reasonable accuracy. Over the years, a number of machine learning models have been trained for various important human targets. However, machine learning models are limited by the quality of the data used to train them. In this work, the focus is on important toxicity endpoints that are being evaluated by next-generation risk assessments, including developmental and reproductive toxicity. Chapter 3 of this work investigates the use of Tanimoto similarity to determine the suitability of using transfer learning. It was found that when the predicted test accuracy (P) or the average similarity between datasets (S) is 70% or more, the machine learning model is likely to have high test accuracy when predicting on the test dataset. In Chapter 4, the creation of a new database for developmental and reproductive toxicity allows for newer machine learning models to be trained whose performances have been reported in this work. Models with about 68% accuracy for developmental toxicity and 80% for reproductive toxicity were trained. The suitability of transfer learning using the models for the two toxicity endpoints has also been investigated and several receptor bindings have been identified as possible mechanisms leading to developmental toxicity or reproductive toxicity.

Preface

As I am writing this, I look back fondly on the memories of the journey thus far. Being an international student, it can be said that one is alone in a foreign land. Fortunately, the support provided along the way has made it more manageable.

My supervisor, Jonathan, could always be counted on for timely help and assistance when it was required. His always optimistic approach to life and problems have brightened up rainy days more than once. I sincerely thank him for all the help he has provided on this challenging journey.

I would also like to extend my thanks to my colleagues who were very helpful with their insights and discussions, in particular Dr Timothy E. H. Allen for his useful criticisms and discussion about part of Chapter 2 content. In addition, I thank Dr Katarzyna R. Przybylak who has provided the EPA Integrated Chemical Environment (ICE) data in Chapter 4 as well as for her assistance and support.

I also thank my family, who have always been supportive of this pursuit. Without them, I would not have made it this far. Just knowing that they are supporting me keeps me motivated when things do not go smoothly.

Table of Contents

List of Abbreviations

Abbreviation	Definition
ACC	Accuracy
ANN	Artificial neural network
AOP	Adverse outcome pathway
AUC	Area under the receiver operating characteristic curve
AutoML	Automated machine learning
CNN	Convolutional neural network
DART	Developmental and reproductive toxicology
DF	Difference between the predicted and actual ML model test accuracy
DILI	Drug-induced liver injury
DNN	Deep neural network
DT	Decision tree
ECFP	Extended-connectivity fingerprint
ECHA	European Chemicals Agency
EPA	Environmental Protection Agency
FN	False negative
FP	False positive
GEPSVM	Generalised eigenvalue proximal support vector machine
GPU	Graphical processing unit
ICE	Integrated chemical environment
IVIVE	*in vitro* assay data towards *in vivo* data
kNN	k-nearest neighbour
LASSO	Least absolute shrinkage and selection operator
LMO-CV	Leave-many-out cross-validation
LOO	Leave-one-out
LSTM	Long short-term memory

MACCS	Molecular ACCess System
MCC	Matthews correlation coefficient
MCCV	Monte Carlo cross-validation
MIE	Molecular initiating event
ML	Machine learning
NB	Naive Bayes
NGRA	Next Generation Risk Assessment
NN	Neural network
OECD	Organisation for Economic Co-operation and Development
P	Predicted test accuracy
QPP	Quadratic programming problem
QSAR	Quantitative structure-activity relationship
RMSE	Root-mean-square error
RF	Random forest
S	Average similarity between datasets
SE	Sensitivity
SP	Specificity
SVM	Support vector machine
SVR	Support vector regression
TN	True negative
TP	True positive
TWSVM	Twin support vector machine
US	United States
USA	United States of America

Chapter 1: Thesis overview

This thesis is about machine learning in predictive toxicology, with a focus on developmental and reproductive toxicity as well as transfer learning. In this chapter, an outline of the entire thesis is presented.

Chapter 2 is a literature review of the recent background of machine learning in predictive toxicology, and also covers several key terminology and concepts used in the remainder of the work. For predictive toxicology, the reliable prediction of the toxicity of a compound has always been a major concern in risk assessments of new products or drugs. With animal testing being phased out, there is an urgent need for alternative methods for determining toxicity. Machine learning, an in silico method, is one such popular choice given that the results can be obtained quickly with reasonable accuracy. Over the years, a number of machine learning models have been trained for various important human targets. Chapter 2 summarises and compares the machine learning models used in predictive toxicology as well as describing the current difficulties of using machine learning in predictive toxicology.

Chapter 3 of this work investigates the use of Tanimoto similarity to determine the suitability of using transfer learning between datasets of human targets such as AChE, ADORA2A, AR, hERG, SERT etc.. A total of 79 datasets for human targets were investigated in Chapter 3. A metric (S) using molecular fingerprints as features and the Tanimoto similarity between molecules which is a measure of the average similarity between two datasets was developed. This metric can be used to measure how similar two datasets are, even if the datasets are from different human targets. This has implications for the transferability of a machine learning model trained on the first dataset and used to predict the properties of molecules in the second dataset. It also represents a measure of the relationship between the human targets being investigated. It was found that when the predicted test accuracy (P) or the average similarity between datasets (S) is 70% or more, the machine learning model is likely to have high test accuracy when predicting on the test dataset. This allows for the possibility of using data from other targets to supplement a lack of data when using machine learning for modelling.

In Chapter 4, a closer look is taken at modelling the binary classification of the general developmental toxicity and reproductive toxicity (DART) endpoint which is known to be highly complex. The largest known database for the prediction of the general DART endpoint containing 3245 compounds (1662 positives, 1583 negatives) has been newly constructed toaid

in this study. Compared to using older data, the new database offers users a more updated set of chemicals as well as providing a list of data sources in one place for ease of use. The new database allows for newer machine learning models for DART to be trained and their performances have been reported in this work. Best-performing models with about 68% accuracy for developmental toxicity and 80% for reproductive toxicity were trained. A consensus approach was also adopted to assist users in analysing if a prediction should be trusted or otherwise. The use of transfer learning also shows comparable results to models trained from scratch, indicating that there is reasonable overlap between developmental toxicity and reproductive toxicity. Finally, the predictions made by the models reported in this work can be applied for screening purposes or in next generation risk assessment frameworks where they could complement other methods in the protection of human health.

Finally, Chapter 5 concludes this work and summarises the key findings demonstrated in this thesis. This chapter also gives some suggestions to further improve on the work done in this thesis as well as providing an outlook for the future.

Chapter 2: Machine learning in predictive toxicology; recent applications and future directions for classification models[1]

2.1 Introduction

Machine learning is a recent field that has advanced computational chemistry with numerous applications such as drug discovery, cheminformatics, and predictive toxicology.[2-22] Machine learning generally involves building a model, training the model, performing validation, repeating training and validation until a suitable model is obtained, and finally testing the model on data not previously exposed to the model (Figure 1). The goal of a machine learning model is to pick out patterns from the input data, or generalise these data, and apply the results to unknown test data.[4-6, 16-20, 22]

These models can be used for predictive toxicology, which is a field that revolves around *in silico* predictions of *in vivo* toxicological effects, such as for drug candidates or drugs,[6, 10, 11, 16, 23-27] consumer products,[28] agrochemicals,[29] and foods.[30,31] Historically, the toxicities of new chemicals were determined through *in vivo* studies, pre-clinical trials, and clinical trials.[23-27, 32, 33] The toxicity of a drug or drug candidate should be determined before it reaches the market or before any clinical trials are performed, where there is a risk of causing severe adverse effects to humans.[23, 24, 26, 27, 32-34] However, determining the toxicity of new compounds is challenging due to a wide variety of potential metabolic products,[34, 35] idiosyncratic effects which only occur in small parts of the population,[24, 27, 33, 34] and the general complexity of *in vivo* systems. These difficulties in determining toxicity have resulted in drug withdrawals or

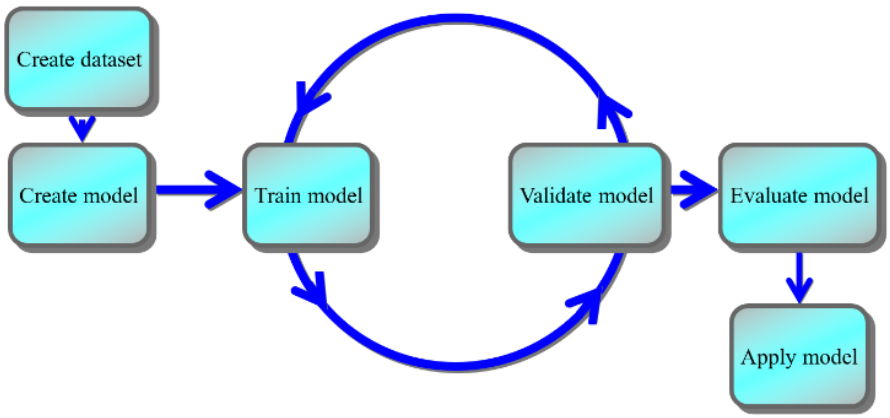


Figure 1. Outline of machine learning.

drug candidates being terminated at the pre-clinical or clinical phase.[23-27, 32-37] Generally, the drug attrition rate is reported to be about 90% to 95% for Phase I trials.[31, 38] Such a high failure rate only contributes to the high cost of drug development, which totalled $2.59 billion in the USA in 2014, and is expected to continue increasing in the future.[39] For these reasons, and due to ethical concerns and regulatory advances,[40-42] there has been a shift towards the use of *in silico* methods for predicting the toxicity of compounds.[31, 43-47]

In silico methods for predictive toxicology include algorithms such as machine learning, quantitative structure-activity relationship (QSAR), expert-based systems, and read-across.[16, 17, 20, 48-50] This thesis chapter focuses on machine learning in predictive toxicology, and cover a variety of toxicity endpoints, of which the main types are hepatotoxicity, carcinogenicity/genotoxicity/mutagenicity, and cardiotoxicity. The focus is on classification algorithms, which treat toxicity as an active/inactive question, rather than on regression models, which attempt to make quantitative predictions of toxicity. Several software packages have been developed for predictive toxicology such as Derek Nexus and the OECD QSAR Toolbox.[51-54] Machine learning models are generally treated as black boxes, and the decisions made by them are sometimes unclear, or at least unexplained. In contrast, some mechanistic QSAR models and expert systems can be understood and interpreted, although this is not true for all of them. Machine learning also has the advantage that it is often able to handle complex problems more effectively and scales well to many different tasks, as evidenced by the successes so far in predictive toxicology.[6, 17-19, 21, 22, 55-57]

Some people consider machine learning algorithms as a type of QSAR, and some consider it to be different. In general, QSAR modelling refers to using a structure-activity relationship to model a qualitative or quantitative prediction of a label. On the other hand, machine learning refers to using a statistical technique to generalise the data, and obtaining predictions based on the model. In machine learning, structure-activity relationships can be used to model the data, which might give rise to confusion between the two types of modelling techniques. It is also noted that with machine learning, there are other features that can be used to model the data, which need not be the molecular structure.

In predictive toxicology, common databases used to build datasets for machine learning models include ChEMBL, ToxCast, and PubChem.[58-65] These databases contain data for different groups of compounds. For example, ChEMBL has data for drug-like compounds while

ToxCast focuses more on industrial chemicals.[60-64] Other sources of data include results provided by pharmaceutical companies, and data that can be extracted from published papers or the public domain.[66-72] The data are subsequently processed for suitability as model input. For example, missing, invalid, or unnecessary data would usually be removed. Another part of data processing is related to working with imbalanced data which is common in machine learning.[71-82] Several methods have been employed to balance the classes in the dataset to train good quality models.[71-82] These methods, which usually involve oversampling, undersampling, or a combination of both are described in the literature.[71-82] The checking and processing of the raw data, structure curation, as well as checking the quality of endpoint data, are essential steps that are often overlooked, which can result in poor models being developed.[83, 84] If the machine learning model does not incorporate some form of interpretability in its architecture, these poor data could give rise to erroneous model performance.

Once the dataset has been prepared, the next step is to build a machine learning model. Many papers have been published on the models used in machine learning and its use in predictive toxicology.[4, 11, 14, 16-22, 48, 55, 85-88] In general, these models can be classified into three categories: supervised learning, unsupervised learning, and semi-supervised learning (Figure 2).

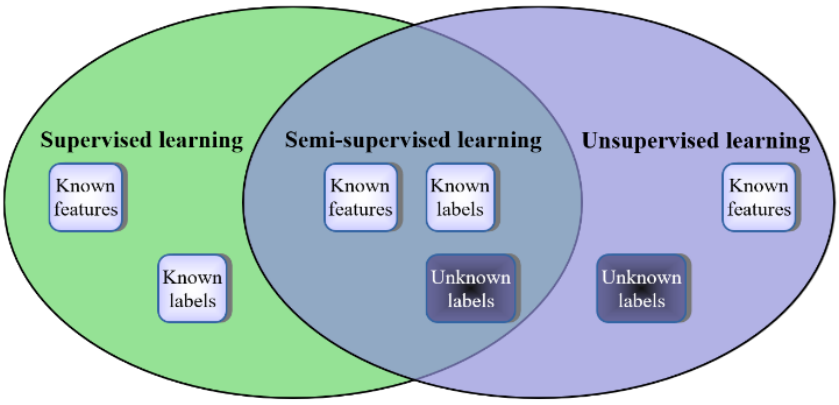


Figure 2. Different categories of machine learning models.

Supervised learning refers to models that are trained on a dataset containing known inputs (features) and outputs (labels).[89-93] The model predicts the labels while the accuracy is defined as the difference between the predicted labels and the experimental labels. These models are usually used for classification purposes. Unsupervised learning refers to models that are trained on a dataset containing known features but unknown labels, and this is commonly used for clustering or pattern recognition.[9, 91, 94-96] Lastly, semi-supervised learning aims to improve

model performance compared to the former two model categories by making use of both labelled and unlabelled data.[93, 97-104]

In predictive toxicology, supervised learning is commonly used as it can classify the input data into different classes (binary or multi-class classification) or labels (multi-label classification).[20, 47, 55, 93, 105, 106] For example, a supervised learning model can be used to predict reproductive toxicity given the molecular fingerprint of a compound.[107] Supervised learning can also be used for regression-based tasks, such as for predicting the quantitative value associated with compound toxicities.[108] In contrast, unsupervised learning and semi-supervised learning are less commonly used in predictive toxicology.

The thesis chapter focuses on supervised machine learning in predictive toxicology, with several model types that have been used in predictive toxicology being analyzed. These model types will be introduced in order of increasing complexity:

1. Regression models
2. k-nearest neighbours (kNNs)
3. Decision trees (DTs)
4. Naive Bayes (NB)
5. Support vector machines (SVMs)
6. Random forest (RF) and ensemble learning
7. Neural networks (artificial neural networks (ANNs), deep neural networks (DNNs), Convolutional neural networks (CNNs))

In this chapter, I look at the overall situation, as well as a more focused analysis on some toxicity endpoints, building on earlier reviews on machine learning models in predictive toxicology.[13, 19, 57]

In predictive toxicology, chemical structures are usually represented by features that can be processed by machine learning methods.[55] This can either take the form of molecular descriptors, molecular fingerprints, or both.[55] Molecular descriptors include features such as atom count, logP (the logarithm of the partitioning coefficient between n-octanol and water, used as a quantitative measure of lipophilicity), solubility etc., and are commonly obtained using cheminformatics toolkits.[55] Molecular fingerprints represent the molecule as a binary string, with each bit in the string corresponding to the fragments or substructures in the

molecule and the bit value representing the absence or presence of that fragment or substructure.[55] More detailed explanations of this subject can be found in the literature.[18, 21, 55, 79, 95, 109-126] In predictive toxicology, commonly used molecular fingerprints include MACCS (Molecular ACCess System) and extended-connectivity fingerprints (ECFPs) such as Morgan fingerprints.[110, 127-129]

Finally, to determine the reliability of the results and the quality of these models, validation methods such as hold-out validation, k-fold cross-validation, or leave-one-out cross-validation are used.[6, 13, 19, 122-124, 130-136] Using validation methods to test the models allows for the assessment of the models' robustness and the reliability of the results obtained. The choice of validation method varies depending on the individual and task due to their inherent advantages and disadvantages. Another part of machine learning: performance measures and metrics have also been reviewed in the literature.[21, 55, 73, 105, 125, 137, 138]

2.2 Validation methods

2.2.1 Hold-out validation

Hold-out validation refers to a method where the dataset is split into a training and test set (Figure 3A).[122, 124] The data which has been partitioned into the training set is subsequently used for training the model and it is validated by calculating performance statistics using the test set. In the case where the test set is independent of the data used for building the model, this is known as external or independent validation. Generally, the test set contains around 20% of the total dataset. In some cases, the data are split into three partitions to prevent hyperparameter bias. The training set is used for training, the validation set for hyperparameter tuning and the test set for final performance review.

2.2.2 k-Fold cross-validation

k-fold cross-validation refers to a method where the dataset is split into k groups (Figure 3B), with k being chosen based on the dataset.[123, 124, 130, 133, 136] Commonly used values of k include 5 or 10. One of the groups is set aside as the test set while the model is trained on the remaining groups. This process is repeated iteratively until all groups have been chosen to be the test set once. Model performance is taken to be the average of the test set performance over all groups.

2.2.3 Leave-one-out cross-validation

Leave-one-out (LOO) cross-validation is a special case of k-fold cross-validation, where the number of groups equals the number of data points (Figure 3C).[109, 115, 134, 135]

Regarding the model training process and evaluation, LOO cross-validation follows the process as described earlier in k-fold cross-validation. Model performance using this validation method is similarly taken to be the average across all runs. Generally, LOO cross-validation is the most computationally expensive due to the large number of training cycles required, while the model also tends to overestimate performance. LOO cross-validation is best used for small datasets to offset these disadvantages.

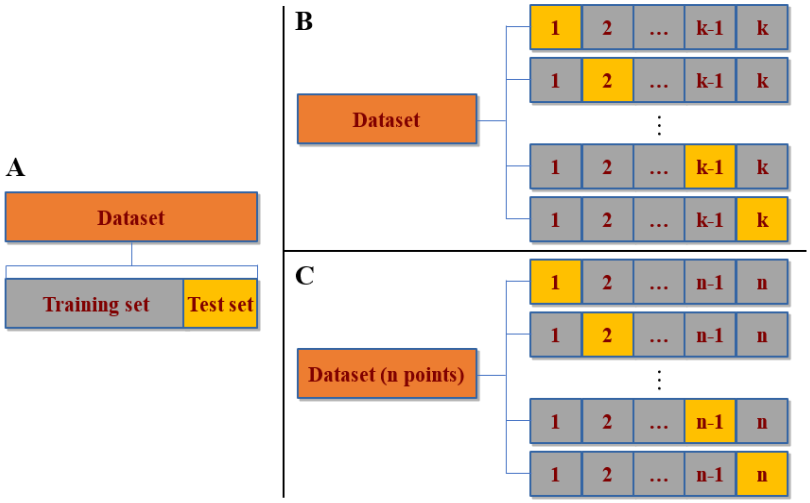


Figure 3: Common validation methods used in machine learning. (A) Hold-out validation, (B) k-fold cross-validation, (C) Leave-one-out cross-validation.

2.2.4 Leave-many-out cross-validation

Leave-many-out cross-validation (LMO-CV), also known as Monte Carlo resampling, or bootstrapping, is another cross-validation technique that is used in the field of machine learning.[139, 140] It involves leaving out all possible subsets of m examples of the data.[140] In a similar fashion to LOO cross-validation, in LMO-CV the training data are split into two subsets: a subset of m examples which is used for validation and a subset of (n-m) examples for training the model.[140] In total, there are C_n^m splits that can be carried out on the training examples.[140] LMO-CV carries out this procedure on all possible C_n^m cases, resulting in an exhaustive procedure that is computationally expensive.[140]

2.2.5 Monte Carlo cross-validation

Another validation technique used in machine learning is Monte Carlo cross-validation (MCCV). This method randomly split the samples into two parts $S_c(i)$ (of size n_c) and $S_v(i)$ (of size n_v), and this procedure is repeated N times (i = 1, 2, …, N), where n_c and n_v are the number of samples in the calibration set and validation set respectively.[141] This is defined in equation (1):[141]

$$MCCV_{nv}(k) = 1Nn_v \sum i = 1N\left\|y_{S_v(i)} - \hat{y}_{S_v(i)}\right\|^2 \quad (1)$$

MCCV reduces the computational complexity drastically because of the reduction in the number of samples.[141] In general, $N = n^2$ is sufficient for $MCCV_{nv}$ to perform as well as CV_{nv} (cross-validation).[141] As compared to LOO cross-validation, MCCV has a larger probability to choose the correct number of components in a model.[141] For the dataset, in order to obtain a larger probability using MCCV, the lesser the number of samples is required for validation.[141]

2.3 Search protocol

All references for the machine learning models (Table 3) were obtained by searching the first 50 pages of results in Google Scholar using the model type, the keywords "machine learning" and "toxicity", and the default settings, with all results being sorted by relevance. Each hit was manually checked to ensure relevance to the topic. The year range was restricted to 2010-2020 to obtain the most recent developments. QSARs, expert-based systems, read-across methods, and preprints were excluded from the results. The most recent search was conducted in March 2020.

A similar search was performed with the Web of Knowledge (Web of Science Core Collection) using the same keywords and the default settings. The results from the Web of Knowledge were refined with the same criteria used with Google Scholar. It was found that there were fewer hits from the Web of Knowledge compared to the search performed using Google Scholar with the hits from Web of Knowledge generally being included in the search from Google Scholar (Table 1). Some of the hits from the Web of Knowledge are not included in the hits from the first 50 pages Google Scholar, perhaps due to the references lacking the keywords specified during the search, or the ordering of the hits in Google Scholar which caused some of the hits to be outside the first 50 pages. Additionally, the references found using Google Scholar were also searched in the Web of Knowledge using the title as the search criteria and the results are recorded in Table 1. In order to give a more complete picture of the

developments in the field, the hits from the Web of Knowledge are also included in the analysis. This search protocol generated a total of 43 references which are listed in Table 3.[10, 18, 47, 142-181]

Table 1. Comparison of literature searches in Google Scholar and Web of Knowledge for the year range 2010-2020.

Machine learning method	Total number of papers from both databases*	Number of papers found during Google Scholar search	Number of papers found during Web of Knowledge search	Number of identical papers found in both databases' search	Number of papers from Google Scholar found in Web of Knowledge
Regression models	2	2	1	1	2
kNNs	8	6	4	2	6
Decision trees	5	2	4	1	1
NB	14	9	5	0	9
SVMs	15	10	8	3	10
Ensemble learning and RF	13	9	9	5	9
Neural networks	13	10	5	2	9
Total	70	48	36	14	46

*Papers with multiple machine learning methods are counted as belonging to the groups they appear in when they are searched. This means that the papers can be counted multiple times and thus the overall total in the table includes this result.

In order to find out if there has been a shift in the situation since 2000-2009, papers were searched using the same criteria as Table 1 while restricting the year range to 2000-2009 with the results being presented in Table 2. However, to speed up the search, only the titles and the abstracts of the papers were considered during the searching process unlike the results in Table 1 where the results were verified manually. This is in line with the intention to estimate the situation for the year range 2000-2009.

Table 2. Comparison of literature searches in Google Scholar and Web of Knowledge for the year range 2000-2009.

Machine learning method	Total number of papers from both databases*	Number of papers found during Google Scholar search	Number of papers found during Web of Knowledge search	Number of identical papers found in both databases' search	Number of papers from Google Scholar found in Web of Knowledge
Regression models	6	5	1	0	5
kNNs	3	2	1	0	2
Decision trees	3	3	1	1	3
NB	3	2	1	0	2
SVMs	9	5	4	0	4
Ensemble learning and RF	3	2	1	0	2
Neural networks	8	6	3	1	5
Total	35	25	12	2	23

*Papers with multiple machine learning methods are counted as belonging to the groups they appear in when they are searched. This means that the papers can be counted multiple times and thus the overall total in the table includes this result.

It was also generally observed that the total number of results using Google Scholar for 2000-2009 is significantly less than 2010-2020 (eg. ca. <1000 compared to >10000 for SVM, 1610 vs. 15600 for neural networks). Therefore, it can be concluded that there has been a general increase in the number of papers published for the field of predictive toxicology for the year range 2010-2020. This can also be seen in Figure 4 which shows the changes in the number of papers over the year ranges used for the search criteria.

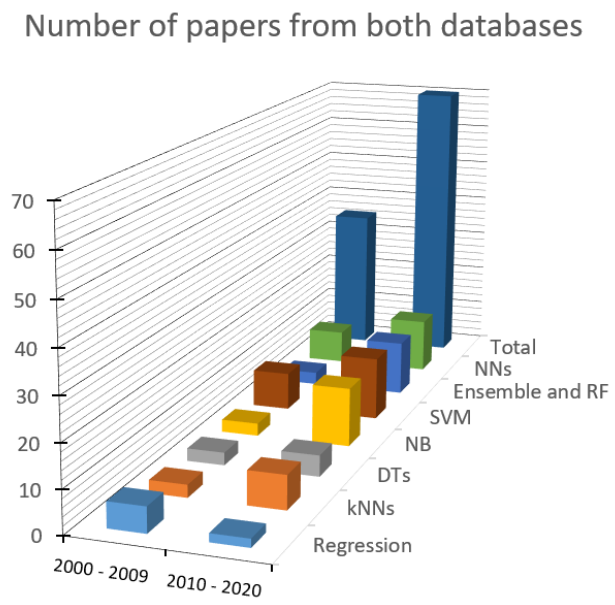


Figure 4. Comparison of results obtained using the search protocol over the year ranges specified.

However, compiling an exhaustive list of relevant articles on this topic when using this search criteria has limitations. This is because articles can list the names of machine learning methods used without explicitly indicating that they represent machine learning and vice versa. For example, it was observed during the search process that titles of papers generally do not include the machine learning method used *i.e.* They just specify "machine learning" while the machine learning method used is usually specified in the abstract of the paper. A similar situation would also apply for the toxicity endpoints and terms. Such limitations would result in papers being missed or papers being counted wrongly if they are not checked manually for relevance, which is time-consuming. Therefore, to supplement the results of the search criteria obtained in Table 1 and Table 2, several papers obtained using a manual search which have a significant impact in the field of predictive toxicology have been chosen to be highlighted.

In particular, the Tox21 program, which was developed in accordance with the National Research Council's vision of testing toxicity in the 21st century.[182, 183] The large amount of publicly accessible data generated due to the collaboration of many institutions have promoted the development and use of high-throughput screening assays.[182, 183] In this chapter, Table 3 shows the results obtained by the search which generally involves predicting the toxicity of drug or drug-like compounds; in other words, pharmaceuticals. As compared to pharmaceuticals which is the focus in recent times (Table 3), Tox21/ToxCast is different

because the focus is on environmental chemicals.[182, 183] Several successful machine learning models have also been built using Tox21/ToxCast data, which have demonstrated the potential of these data in predictive toxicology.[143, 184-188]

One paper to be highlighted is ToxPrint which is a widely used set of structural features from molecules in toxicity databases that can be represented as chemical fingerprints.[144, 189] This was developed by Yang et al in 2015 and is based on various toxicity prediction models and safety assessment guidelines by several institutions including the Food and Drug Administration.[144, 189] ToxPrint has also begun to be used with machine learning models, including a recent study predicting estrogen binding.[190]

Another paper to be highlighted involves XGBoost. XGBoost is a scalable tree boosting method used in machine learning, and has received recognition in numerous data mining and machine learning challenges.[191] XGBoost is highly scalable in all scenarios which has contributed to its success in machine learning.[191] Further details can be found in the work by Chen et al 2016.[191]

Table 3: A summary of the results of all machine learning methods.

No.	Machine learning method	Toxicity endpoint	Accuracy (%)	AUC (%)	SE (%)	SP (%)	Validation type	Dataset size	Reference Number
1	Regression (Linear)	Cardiotoxicity	-	75	-	-	10-fold	1917	[162]
2	Regression (Ridge)	Cardiotoxicity	-	77	-	-	10-fold	1917	[162]
3	Regression (Partial least square)	Molecular toxicity	82	-	-	-	Hold-out	2849	[173]
4	kNN	Aquatic toxicity	84	92	84	85	10-fold	1005	[177]
5	kNN	Carcinogenicity	-	-	84	84	External	661	[178]
6	kNN	Cardiotoxicity	82	78	82	57	External	206	[179]
7	kNN	Genotoxicity	86	93	80	90	5-fold	576	[180]
8	kNN	Hepatotoxicity	62	52	91	20	External	978	[181]
9	kNN	Hepatotoxicity	78	-	79	76	10-fold	1274	[142]
10	kNN	Hepatotoxicity	-	-	62 ± 20	92 ± 14	10-fold	288	[143]
11	kNN	Organ toxicity	-	-	92 ± 8	78 ± 6	5-fold	-	[144]

12	DT	Food-related toxicity	0.23[+]	-	-	-	10-fold	94	[175]
13	DT	Genotoxicity	82	81	75	87	5-fold	576	[180]
14	DT	Hepatotoxicity	89*	-	-	-	10-fold	575	[176]
15	DT	Hepatotoxicity	-	-	74 ± 16	94 ± 5	10-fold	288	[143]
16	DT	Organ toxicity	-	-	89 ± 12	76 ± 7	5-fold	-	[144]
17	NB	Aquatic toxicity	77	81	70	84	10-fold	1005	[177]
18	NB	Carcinogenicity	68 ± 2	-	60 ± 8	75 ± 10	5-fold	834	[145]
19	NB	Developmental toxicity	83	-	90	67	5-fold	232	[146]
20	NB	Genotoxicity	85	90	89	81	5-fold	576	[180]
21	NB	Hepatotoxicity	-	-	73	73	5-fold	336	[147]
22	NB	Hepatotoxicity	-	-	70 ± 15	85 ± 7	10-fold	288	[143]
23	NB	Immunotoxicity	-	78	73	70	Hold-out	44615	[148]
24	NB	Mitochondrial toxicity	81 ± 1	-	88 ± 4	77 ± 4	5-fold	226	[151]
25	NB	Mutagenicity	90.9 ± 0.3	-	39 ± 4	95 ± 1	5-fold	5159	[149]
26	NB	Mutagenicity	65	-	-	-	10-fold	3903	[150]
27	NB	Myelotoxicity	82 ± 3	-	76 ± 6	84 ± 3	External	727	[152]
28	NB	Nephrotoxicity	-	-	62	78	Stratified 3-fold	27	[153]
29	NB	Respiratory toxicity	84	-	84	85	5-fold	993	[154]
30	NB	Urinary tract toxicity	84	-	84	85	5-fold	173	[155]
31	SVM	Acute oral toxicity	90	-	-	-	External	8102	[47]
32	SVM	Acute toxicity	70	72	85	59	External	321	[156]
33	SVM	Aquatic toxicity	89	94	89	89	10-fold	1005	[177]
34	SVM	Carcinogenicity	-	-	73	79	External	499	[10]
35	SVM	Carcinogenicity	68 ± 3	73 ± 3	65 ± 5	72 ± 5	5-fold	802	[157]
36	SVM	Carcinogenicity	78	-	84	74	External	661	[178]
37	SVM	Cardiotoxicity	86	72	88	29	External	206	[179]
38	SVM	Cardiotoxicity	87	-	90	74	10-fold	1501	[158]
39	SVM	Genotoxicity	89	95	92	84	5-fold	576	[180]

40	SVM	Hepatotoxicity	-	-	58 ± 16	99 ± 6	10-fold	288	[143]
41	SVM	Hepatotoxicity	75	61	93	38	External	978	[181]
42	SVM	Hepatotoxicity	83	89	93	68	External	1731	[159]
43	SVM	Mutagenicity	72	-	69	74	10-fold	1696	[160]
44	SVM	Nephrotoxicity	-	-	79	84	Stratified 3-fold	27	[153]
45	SVM	Ototoxicity	85	-	82	92	External	536	[161]
46	Ensemble learning/RF	Aquatic toxicity	86	93	83	89	10-fold	1005	[177]
47	Ensemble learning/RF	Carcinogenicity	70 ± 3	77 ± 3	67 ± 5	73 ± 4	5-fold	802	[157]
48	Ensemble learning/RF	Carcinogenicity	74	-	65	80	Hold-out	661	[178]
49	Ensemble learning/RF	Carcinogenicity	68 ± 3	74 ± 3	64 ± 5	73 ± 4	5-fold	802	[157]
50	Ensemble learning/RF	Cardiotoxicity	82	94	65	98	10-fold	2901	[163]
51	Ensemble learning/RF	Cardiotoxicity	97	-	-	-	External	522	[164]
52	Ensemble learning/RF	Genotoxicity	94	96	95	93	External	576	[180]
53	Ensemble learning/RF	Hepatotoxicity	69 ± 3	75 ± 3	76 ± 3	62 ± 5	5-fold	993	[166]
54	Ensemble learning/RF	Hepatotoxicity	74	79	-	-	10-fold	281	[165]
55	Ensemble learning/RF	Hepatotoxicity	-	-	58 ± 15	97 ± 5	10-fold	288	[143]
56	Ensemble learning/RF	Hepatotoxicity	71 ± 3	76 ± 2	80 ± 4	60 ± 5	5-fold	1117	[166]
57	Ensemble learning/RF	Hepatotoxicity	73	-	77	66	10-fold	1274	[142]
58	Ensemble learning/RF	Nephrotoxicity	-	-	89	75	10-fold	30	[167]
59	NN (ANN)	Acute toxicity	-	70	100	53	External	321	[156]
60	NN (ANN)	Aquatic toxicity	88	94	87	89	10-fold	1005	[177]
61	NN (ANN)	Genotoxicity	87	94	91	82	5-fold	576	[180]
62	NN (ANN)	Hepatotoxicity	82	-	71	98	External	475	[169]
63	NN (ANN)	Mutagenicity	60	-	40	81	10-fold	1696	[160]
64	NN (ANN)	Mutagenicity	80	87	84	75	5-fold	6094	[170]
65	NN (DNN)	Cardiotoxicity	93	97	93	91	Hold-out	3954	[168]
66	NN (DNN)	Cardiotoxicity	98	-	-	-	External	522	[164]
67	NN (DNN)	General	-	84	-	-	External	11764	[18]

68	NN (DNN)	Hepatotoxicity	81	-	82	80	External	475	[169]
69	NN (Graph CNN)	Cardiotoxicity	-	96	-	-	-	3954	[168]
70	NN (CNN)	General	-	78	-	100	-	10588	[171]
71	NN (CNN)	General	-	85	-	-	-	7438	[172]
72	NN (CNN)	Hepatotoxicity	63	62	64	62	-	7630	[174]

*corrected classification rate (CCR) used instead of accuracy.

[+]Error between the predicted value vs. the actual value was used instead of accuracy.

2.4 Model types

2.4.1 Regression models

A regression task involves predicting a numerical response variable using several predictor variables by learning a model that minimises the loss function.[192] First, a distinction is made between regression models and support vector regression (SVR), which is an application of support vector machines (SVM), and will be covered in a later section of this chapter. Also, in the literature, kernel functions are sometimes referred to as regression models. In this chapter, a distinction is made between the two: the term kernel functions will be reserved for the SVMs while regression models will be discussed in this section.

Regression models, which are statistical models, can be broadly classified as linear regression and non-linear regression. These include linear, multivariate linear, polynomial, stepwise, ridge, and least absolute shrinkage and selection operator (LASSO).[193-208] These regression models are used for quantitative predictions unlike the standard classification models. Examples of these quantitative characteristics of toxicity include LD_{50}, LC_{50}, IC_{50}, and EC_{50}. While linear regression models have low computational cost, their linear nature limits their ability to model complex problems, unlike non-linear regression models.[87, 209, 210] However, using non-linear regression models will increase the computational cost.

In predictive toxicology, regression functions have been employed in several works. These models, as well as all machine learning methods in this chapter, will be measured by performance metrics which include accuracy (Q), sensitivity (SE), specificity (SP), and area under the receiver operating characteristic curve (AUC). Even though the accuracy was chosen as the performance metric for regression model types, it is acknowledged that R^2 and root-mean-square error (RMSE) are more adequate performance metrics to gauge the quality of the model. However, in line with an intention to give an estimate of the performance of regression

models as compared to the other machine learning methods, accuracy which is a common performance metric used in machine learning was used. Table 3, entries 1 – 3, summarizes the recent performance of several regression models across different toxicity endpoints.

2.4.2 k-Nearest neighbours (kNNs)

kNN is a non-parametric classifier where the test sample is assigned a class label based on the most frequently occurring class label among the k-nearest neighbours.[211-213] A proximity measure, such as Euclidean distance or Manhattan distance, is used to define the k-nearest neighbours to each test sample.[211-213] All samples are represented by points in an n-dimensional feature space, while the neighbours are taken from a set of objects for which the correct classification or value is known.[212]

More formally, the k-nearest neighbours algorithm is defined as follows: Given a collection of incomplete/unlabelled test data $\{(x_i, y_i), i = n + 1, \dots, n + m\}$, the problem amounts to predicting the class labels for $y^* = \{y_{n+1}, \dots, y_{n+m}\}$ with corresponding feature vectors $x^* = \{x_{n+1}, \dots, x_{n+m}\}$.[211, 214] Thus, the k-nearest neighbours algorithm amounts to classifying an unlabelled y_{n+1} as the most common class among the k-nearest neighbours of x_{n+1} in the training set $\{(x_i, y_i), i = 1, \dots, n\}$.[211, 212, 214] In the algorithm, the value of k is typically a positive integer, usually small (such as $k = 1$), or is chosen based on leave-one-out cross-validation.[211, 212, 214, 215]

If the value of k chosen is too small, it might result in overfitting while if the value of k is too large, it might result in misclassification.[212] While kNN is easy to implement and often gives good performance, it is heavily dependent on the classification accuracy of the test class labels as well as the value of k.[214] Samsudin et al has also outlined several improvements to kNN in the literature.[213] Table 3, entries 4 – 11, summarizes some of the results achieved by kNN models.

2.4.3 Decision trees (DT)

A decision tree is a tree-structured classifier consisting of a root, nodes, and leaves where each node has only one unique path from the root. The decision that the classifier makes at each node is based on decision rules, which depends on the features of the data used. Several papers have explained DTs in great detail and will not be reproduced here.[216-218] Furthermore, in a

later section, random forest which is an ensemble of decision trees will also introduce the basics of decision trees.

DTs have the advantage of model interpretability because the decision rules can be retrieved from the model for each result, unlike complex models like neural networks where each node is based on all of the nodes in the previous layer. However, due to the design of the tree, it is easier for errors to accumulate at each level, and thus a compromise on accuracy and efficiency has to be reached.

In predictive toxicology, decision trees are not commonly used as evidenced by the data from Table 1 and Table 2. This could be because the toxicity endpoints are complex and thus a simple model cannot generalise all the patterns in the data. Table 3, entries 12 – 16, summarizes some of the results achieved by DT models.

2.4.4 Naive Bayes (NB)

In Bayesian classification, the given data are hypothesized to belong to a particular class.[219] The probability that the hypothesis is true is then calculated.[219] Another way to describe the Bayesian classifier is shown in equation (2), where the Bayesian classifier is defined as obtaining the posterior probability $P(C_i \mid A_1, \ldots, A_n)$ of each class C_i, using Bayes rule.[220, 221]

$$P(C_i \mid A_1, \ldots, A_n) = P(C_i)P(A_1 \mid C_i) \ldots P(A_n \mid C_i)/P(A) \quad (2)$$

This equation makes the simplifying assumption that given the class, the attributes, A, are independent and thus the likelihood can be obtained by the product of the individual conditional probabilities of each attribute.[220, 221] This is called a naive Bayes (NB) classifier.

Naive Bayes models are efficient, generally robust and have high accuracy.[220, 221] However, the NB model classification accuracy decreases when the attributes are not independent.[220] Also, NB models cannot deal with non-parametric continuous attributes.[220] Some improvements such as feature selection have been carried out to tackle these issues.[222] Additional details about these improvements can be found in the literature.[220, 222-225] Table 3, entries 17 – 30, shows the results obtained for NB models in predictive toxicology.

2.4.5 Support vector machines (SVM)

The next model type to be introduced is SVM, which was introduced by Vapnik et al.[226] It is based on the structural risk minimization principle which was developed from statistical learning theory.[226-230] Vapnik et al state that the basis of the principle is to control the Vapnik–Chervonenkis dimension (VC-dimension) to minimise the guaranteed risk, which is the sum of the empirical risk and the confidence interval.[227] This, however, involves a trade-off as the minimum empirical risk decreases while the confidence interval increases as the VC-dimension increases.[227] Yan et al describes this in another way, which is that SVMs, which are maximum margin classifiers, simultaneously minimize the empirical classification error and maximize the geometric margin.[228]

According to Vapnik et al, an SVM maps the input vectors into some high dimensional feature space Z through some non-linear mapping chosen *a priori*.[226] This allows a hyperplane to be constructed between the data points, where there is a margin of separation between the two classes.[226] If the training data cannot be separated with error, the algorithm would separate the training set with a minimal number of errors, which results in a soft margin SVM.[226]

SVMs usually results in a feature space where the data are linearly separable, even when the initial feature space is non-linear. In the case of a linear classifier, the feature space is separated by a hyperplane with equation (3), where w is the weight vector, x is the input vector, and b is the bias (Figure 5).[21, 226, 231-233]

$$w^T \cdot x + b = 0 \tag{3}$$

The geometric margin is thus represented by the constraints shown in (4) and this is shown in Figure 5 as the boundary lines running parallel to the hyperplane.[21, 226, 231-233]

$$w^T \cdot x + b \begin{cases} \geq 1 \; when \; y_i = +1 \\ \leq -1 \; when \; y_i = -1 \end{cases} \tag{4}$$

To construct a linear SVM classifier for a non-linear feature space, kernels such as a Gaussian, radial basis function, or polynomial types are used.[10, 160, 229, 234-239] These kernels are functions that work by mapping the input data which is linearly non-separable into a higher dimensional space where the data are linearly separable.[10, 234] A hyperplane can thus be constructed that separates the two classes, resulting in a situation similar to a linear classifier.

It is known that SVMs are good at pattern recognition, can generalise well, and can handle high-dimensional data.[226, 229, 231, 240] However, the drawbacks of SVMs include difficulty in choosing an appropriate kernel and being time-consuming for large datasets.[234, 236, 240] Improvements have been made to counter these drawbacks. For example, SVMs such as the

generalised eigenvalue proximal support vector machine (GEPSVM) and twin support vector machine (TWSVM) have been developed to reduce the time consumed.[241-243] These SVMs work by constructing two non-parallel hyperplanes instead of a single hyperplane, effectively reducing the quadratic programming problem (QPP) required to generate the hyperplane(s) from a single large QPP to two smaller QPPs.[241-243] Also, multiple kernel functions have been developed which can handle complex classification problems better by adapting better to the characteristics of the data.[244-247] In recent years, several SVMs have been used in predictive toxicology. These results are tabulated in Table 3, entries 31 – 45.

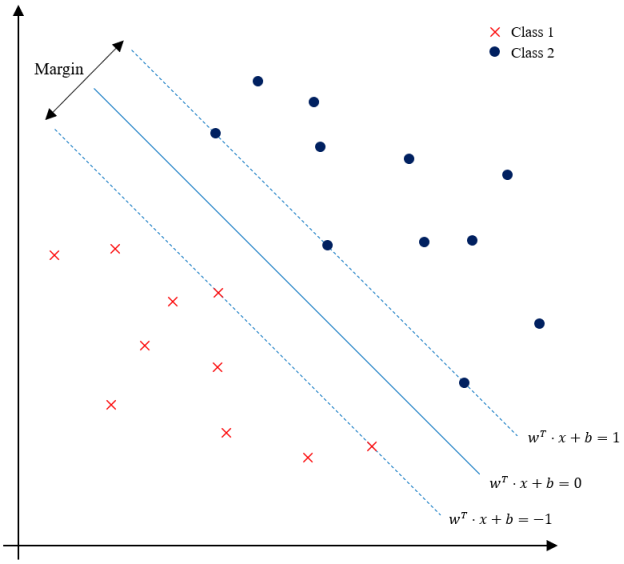


Figure 5. A representation of an SVM with linearly separable classes.

2.4.6 Random forest (RF) and ensemble learning

A review by Sagi et al provides a comprehensive survey of ensemble learning.[248] In this chapter, some of the ideas by Sagi et al will be introduced in this section as a general introduction to this machine learning method. There are also other reviews on ensemble learning.[249-255] In this chapter, a general overview on ensemble learning will be given before focusing on RF as it a popular machine learning method used in predictive toxicology.

Ensemble learning refers to methods that combine multiple inducers to make a decision, typically in supervised machine learning tasks.[248] An inducer, or base-learner, is an algorithm that takes a set of labelled examples as input and produces a model that generalises these examples.[248] By combining multiple models, the error of a single inducer will likely be compensated by other inducers, which leads to better overall performance as compared to a single inducer.[248] In other words, the predictive performance of an ensemble learning model cannot be lower than the predictive performance of each model making up the ensemble.[248]

More formally, the ensemble learning method can be represented as follows: Given a dataset of n examples and m features, $D = \{(x_i, y_i)\}(|D| = n,\ x_i \in R^m, y_i \in R)$, an ensemble learning model φ uses an aggregation function G that aggregates K inducers, $\{f_1, f_2, \ldots, f_k\}$ towards predicting a single output.[248] This is represented by equation (5).[248]

$$\hat{y_i} = \varphi(x_i) = G(f_1, f_2, \ldots, f_k) \tag{5}$$

In an ensemble method, there exist two main types of frameworks.[248] The dependent framework is one where the inducer's output affects the construction of the next inducer.[248] In contrast, the independent framework each inducer is built independently from other inducers.[248] For instance, popular ensemble methods such as AdaBoost, bagging, and random forest (RF) are examples of dependent, dependent, and independent frameworks respectively.[248]

Ensemble learning methods are generally able to handle class imbalance, concept drift, and the curse of dimensionality better as compared to their machine learning counterparts.[248] Also, these methods tend to avoid overfitting as different hypotheses are averaged to give the final result.[248] Moreover, ensemble learning methods decrease the risk of finding a local minimum while also giving a better representation of the data.[248]

However, even with these numerous advantages, there are some considerations when building an ensemble learning model. This includes the individual method's suitability towards the data, difficulty in interpreting the output of the ensemble learning model, the software availability, and the usability, and, sometimes, computational cost.[248] As ensemble learning methods are built up of multiple models, it follows that the total computational cost will at least be equivalent to running each model separately. It is thus likely that significant computational resources need to be allocated to train an ensemble learning method for results to be obtained

within a reasonable time limit. Improvements to ensemble learning methods are also covered in the review by Sagi et al.[248]

Even though ensemble learning uses a general approach eg. Bagging, the intention is to give an estimate of the performance of models that use these general approaches as compared to concrete machine learning methods like RF. Ensemble learning can be interpreted as a group of algorithms eg. Bagging and a machine learning method, or as a group of machine learning models eg. NB and neural networks. Thus, care should be taken when comparing performance metrics of ensemble learning methods to other machine learning methods. It should also be noted that RF is a special case because it is made up of an ensemble of decision trees and is thus treated as a concrete machine learning method. Next, RF will be covered in more detail while details on the other two methods can be found in the literature.[255-260]

Breiman introduced RF in 2001 as a classifier that is made up of multiple tree-structured classifiers (decision trees) $\{h(x, \theta_k), k = 1, ...\}$ where the $\{\theta_k\}$ are independent identically distributed random vectors, and each tree casts a unit vote for the most popular class at input x.[114, 261] Each RF consists of a root, nodes and leaves, with each split representing two branches at each node. (Figure 6). Each node in an RF represents the decision made by the model which is based on a subset of the available features while the leaves represent the outputs of the RF.[114, 261-263] These outputs, which are predictions of each tree, are combined by taking the most common prediction across all trees.[112, 261, 262]

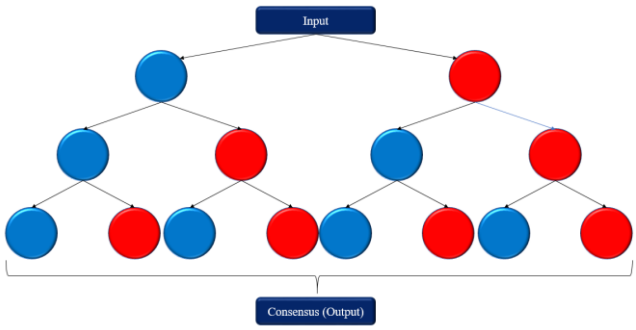


Figure 6. A random forest with two decision trees. The two colours represent the positive or negative decision made at each split.

To build an RF for machine learning, an algorithm has to be chosen. One common algorithm used in literature was developed by Breiman and uses the Bagging principle together with random feature selection.[114, 261] Each tree in the RF is grown by randomly drawing N samples

from the original training set with replacement, also known as bootstrapping which is then used to build the tree.[114, 261] For each node of the tree, a portion of features from the original feature space is randomly drawn without replacement, among which the best split is selected.[114, 261] Throughout the process, no pruning of the tree is performed.[114, 261]

Building an RF also requires that one considers pruning as well as the number of trees. Pruning refers to removing nodes in the RF-based on a criterion which simplifies the RF.[263, 264] On the other hand, the number of trees in an RF affects the generalization of the model.[114, 261] However, it was found that above 10000 trees using Adaboost, or 100 trees for a well-defined value of the number of random features pre-selected in the splitting process using the Forest-RI method, adding more trees does not improve performance.[114, 261]

RFs have several advantages such as they can minimise the issue of overfitting, are resistant to noise with some algorithms, can handle high-dimensional data well, having the ability to ignore irrelevant descriptors, and being interpretable in terms of the decision rules made.[112, 265, 266] RFs are also known to keep the benefits afforded by DTs while achieving better results most of the time.[266] On the contrary, the disadvantages of RF include being susceptible to bias when there are dominant features, as well as placing more emphasis on the correlation among smaller groups of features.[267]

A review by Fawagreh et al in 2014 has covered some of the recent advancements in RF.[268] They mention in the review that the performance of the RF can be improved through the use of different voting methods, or by implementing a weighting scheme for the features or for discarding trees.[268] Another review by Rokach in 2016 on decision forests (including RFs) introduces the models, the methods to construct them, and surveys the state-of-the-art methods in the field.[249] Next, Table 3, entries 46 – 58, summarizes the results of RFs and ensemble learning used in predictive toxicology.

2.4.7 Neural networks (artificial neural networks (ANNs), deep neural networks (DNNs), convolutional neural networks (CNNs))

Neural networks (NNs) can generally be divided into three groups, namely artificial neural networks (ANNs), deep neural networks (DNNs), and convolutional neural networks (CNNs). In this chapter, ANNs and DNNs will be elaborated in more detail first as they as similar, following which CNNs will be explained in more detail.

In recent times, artificial neural networks (ANNs) which excel at pattern recognition and classification have been successfully applied in multiple fields such as novelty detection, renewable energy systems, and image processing.[269-271] The most widely used types of ANNs include feedforward and recurrent neural networks, of which feedforward neural networks are generally enough for most binary classification tasks.[272] In this chapter, the focus will be on feedforward neural networks. Recurrent neural networks such as long short-term memory (LSTM) are not covered in this thesis chapter and can be found in the literature.[273-277] In a typical feedforward neural network, the network is made up of layers, namely the input layer, hidden layers and the output layer. Each layer is in turn made up of independent nodes or neurons, with each node connected to all nodes in the subsequent layer. This is illustrated in Figure 7. In the literature, several strategies for determining the number of nodes as well as hidden layers when building a neural network have been described.[278-280]

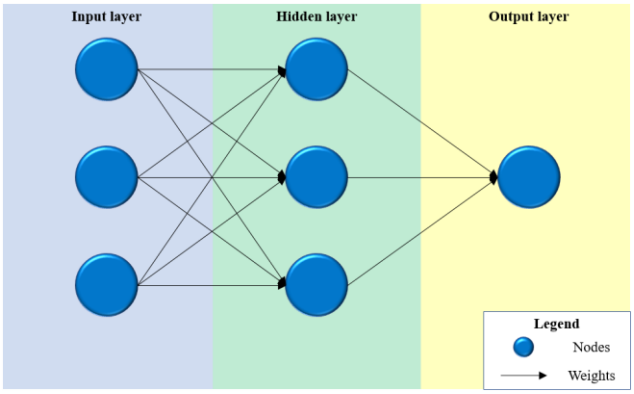


Figure 7. Graphical representation of a feedforward ANN.

In addition to the network architecture, one must choose an activation function which is a function that transforms the activation level of a unit (neuron) into an output signal.[281] Examples of activation functions include linear, sigmoid, Gaussian, and Elliot.[281] A popular and recent activation function called rectified linear unit (ReLU) has been shown to converge quickly for neural networks.[282-284] The ReLU activation function is represented by equation (6).

$$f(u) = max(0, u) \tag{6}$$

Overall, the output in a feedforward neural network can be described by equation (7) where o_i represents the output, σ_i is the output function, b_i is the bias input of the *i*th hidden node, w_{ij} is the weight of the connection from the *j*th input, u_j, to the *i*th hidden node, and M is the total number of inputs.[285]

$$o_i = \sigma_i(b_i + \sum_{j=1}^{M}(w_{ij} * u_j)) \quad (7)$$

The starting weights are usually randomised, with the final weights commonly determined by backpropagation. Backpropagation is a method to calculate the gradient of the error with respect to weights and several algorithms have been developed which include first order and second-order algorithms.[286, 287] One of the commonly used algorithm is the Levenberg-Marquardt (LM) algorithm which is a second-order algorithm.[288] Wilamowski et al describes these backpropagation algorithms in more detail.[288]

Recently, deep neural networks (DNNs) have gained interest due to deep learning, of which there have been several reviews published in the literature.[2, 87, 94, 272, 287, 289-293] These DNNs are neural networks with a deep network architecture, where the number of hidden layers is more than one.[294, 295] By increasing the number of hidden layers, DNNs can handle more complex problems, of which there are successful examples such as skin cancer classification, image classification, and syntheses route planning.[294-296]

Despite the growing popularity of DNNs, one must acknowledge the several limitations of DNNs. These include longer network training times with an increasing number of hidden layers and parameters. Improvements have been made to rectify these issues. For example, graphical processing units (GPUs) have been employed to train neural networks, where the higher processing capabilities of GPUs help reduce training times.[297-300]

On the other hand, ANNs have their own set of limitations which include their expressivity, which shows that they are unable to express certain functions that DNNs are capable of, as well as having lower approximation capability as compared to DNNs.[301] Unfortunately, as these are issues intrinsic to the model type, there are no good solutions for them.

In the field of predictive toxicology, DNNs have been shown to be successful in numerous examples (Table 3). Several papers have also covered on the topic of neural networks or DNNs in the field of predictive toxicology,[11, 17, 105] in particular, the review by Tang et al in 2018.[6] Hence, this thesis chapter will not focus on the recent advances of DNNs, but rather the model performance of DNNs in predictive toxicology. Given the increased performance of DNNs as compared to their ANN counterparts, it is predicted that the use of DNNs in predictive toxicology will become more prevalent. The complexity provided by DNNs could also give

these models an edge when predicting complex toxicity endpoints, for which simpler models might not be able to model well. Nevertheless, the lack of inherent interpretability of the results of DNN models still poses a challenge in the field due to the large number of parameters, weights, and nodes involved.

Convolutional neural networks are similar to ANNs and DNNs in the sense that they also consist of layers and are feedforward networks. However, unlike their counterparts, a CNN typically consists of convolutional and pooling layers stacked on top of each other.[302-304] The fully connected layers that follow these layers interpret the feature representations and perform the function of high-level reasoning, such as classification.[303]

Convolutional layers serve as feature extractors, where the neurons are arranged into feature maps.[305, 306] Each neuron has a receptive field which is connected to the neurons in the previous layer via a set of trainable weights while all neurons within a feature map have weights that are constrained to be equal.[305, 306] Equation (8) shows the representation of the *k*th output feature map Y_k, where x represents the input image, W_k the convolutional filter, the multiplication sign as the 2D convolutional operator used to calculate the inner product of the filer model at each location of the input image, and $f(\cdot)$ the non-linear activation function.[303]

$$Y_k = f(W_k * x) \tag{8}$$

In contrast, the pooling layers reduces the spatial resolution of the feature maps which reduces the number of parameters (controls overfitting), achieving spatial invariance to input distortions and translations.[303] These pooling operations play a role in producing downstream representations that are more robust to the effects of variations in data while still preserving important motifs.[303, 305] Lee et al and Rawat et al describes these pooling operations in more detail.[303, 305] In toxicology, molecules are represented as images, grids, or graphs which are then fed into the CNN for training.[168, 171, 172, 174]

Similar to ANNs and DNNs, CNNs also uses the backpropagation algorithm during training.[303] A typical CNN with both convolutional and pooling layers is shown in Figure 8.

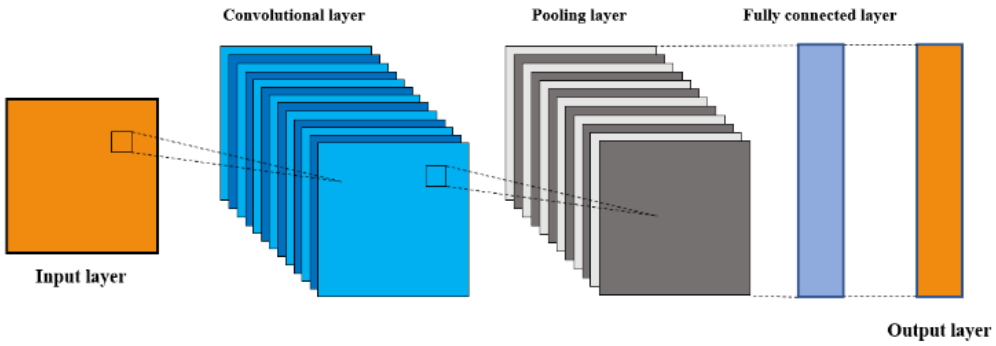


Figure 8. A CNN with convolutional and pooling layers.

CNNs have several advantages over traditional ANNs such as requiring fewer free parameters and being able to deal with the variability of 2D shapes.[303] However, CNNs require a large amount of training data which increases computational cost and lengthens training time.[303] In recent years, new CNNs have been developed such as deep CNNs and graph CNNs which outperforms traditional CNNs. A recent review by Rawat et al covers the topic of deep CNNs more extensively.[303] On the other hand, graph CNNs are the most recent CNNs to be developed which consist of embedding nodes in a graph in Euclidean space.[306] More information about graphs and graph CNNs can be found in the literature.[306-310]

Before the advances of CNNs in recent times, in the field of predictive toxicology, traditional CNNs were uncommonly used. This is because traditional CNNs use images as inputs while in predictive toxicology, molecular structures, fingerprints, or descriptors are more common as inputs. An example of this is the work by Jimenez-Carretero et al[311] In their work, a CNN using cell-based images as features was developed. This is a situation where machine learning is used to generalise the biological properties, rather than the chemical properties which the main focus of predictive toxicology focuses on. In this chapter, the focus is on molecular structure in predictive toxicology as thus far, traditional CNNs have not shown themselves to be useful in this field. Table 3, entries 59 – 72, summarizes some of the results achieved by NN models.

2.5 Overall analysis

The popularity of SVMs and RFs could be attributed to their advantages, such as being efficient while easy to use, as well as being able to generalise the data well. Furthermore, numerous successful models have been built for these two model types, thus contributing to their well-established reputation. On the other hand, recently developed machine learning models, such

as DNNs or ensemble learning, are less popular possibly due to their high computational cost or their complexity, or because they are not as well-established as SVMs and RFs. Machine learning methods perform differently on different datasets and these differences arise from the diverse characteristics of the data, such as the dataset size, class distribution, and the distribution of the data in the feature space.

It is also observed that the simplest machine learning method, regression models are not the most popular model of choice in predictive toxicology. The performance of regression models also falls behind their classification counterparts. However, caution should be taken when comparing the performance of regression models with classification models as the two are inherently different. This illustrates the complexity of the problem of predicting the toxicity of chemicals, where usually, a more complex machine learning model is required to model the problem or the data more effectively. A more complex model might also produce results that one would otherwise miss if a simpler model was used instead, simply because the model does not over-generalise the data. Another probable reason is perhaps the familiarity or ease of use of such machine learning methods, where one is inclined to use a method that is more familiar to raise the chances of building a successful model. Since SVM has been well established in the literature, it continues to be a popular machine learning method for predictive toxicology.

The statistical performance results of all machine learning methods in predictive toxicology covered, including the values for sensitivity (SE), specificity (SP), accuracy, AUC, and the validation type are summarized in Table 3. It is common to see different studies use different performance metrics to measure their model's performance. While the best performance metric to use is still up for debate, such diversity in the performance metrics makes it difficult to compare across different models. Moreover, as the models generally do not use a benchmark dataset, it is once again difficult to compare the performance of different models.

Generally, the machine learning methods in Table 3 have an accuracy or AUC of 75% or above. Those models which reported a lower performance than expected could have experienced issues with the dataset, such as the dataset not being able to generalise well to another test set. Most of the time, it is more likely for there to be a problem with the data or the deployment of the algorithm, rather than the machine learning method of choice which are generally well established.

Based on the data and the search criteria outlined in Table 3, hepatotoxicity, carcinogenicity/ mutagenicity/genotoxicity, and cardiotoxicity are the most common types of toxicity that have been investigated. For this chapter, carcinogenicity, mutagenicity, and genotoxicity have been grouped together as they are quite similar, and this makes it simpler to analyse the results. However, one should note that these are three different endpoints and that usually models for mutagenicity and genotoxicity performed better than for carcinogenicity which is a more complex endpoint (Table 3). Hepatotoxicity is important in general because most toxicity originates from the liver which is the main site of metabolism for drugs; that is hepatotoxic compounds would have adverse side effects *in vivo* and thus are unsuitable to be drugs.[137, 159] Hence, determining the hepatotoxicity potential of a drug candidate would allow for the quick screening of potential drug candidates from all of the compounds. Cardiotoxicity is important because side effects such as cardiac arrest are highly undesirable. Lastly, tests for carcinogenicity and mutagenicity are carried out during drug screening as they are important toxicological endpoints.[228] The Ames test is used as part of a battery of *in vitro* methods which covers different mechanisms that may lead to mutagenicity.[312, 313] *In silico* predictions of the Ames test mutagenicity have also been investigated by Xu et al and Hillebrecht et al, which highlights the need for faster and more accurate predictions of a key test in toxicology.[170, 314] Hillebrecht et al has demonstrated that the expert-based system Derek performs the best for predicting Ames test mutagenicity, though they believe that the fusion of the expert-based system and QSAR techniques would lead to improvements in the predictive power of the *in silico* models.[314] This is also the basis of the FDA and EU guidance on impurity testing in drugs – ICH M7.[315] Therefore, by screening for these major types of toxicity, the number of potential drug candidates can be reduced to a smaller number from a larger range of drug-like compounds.

The distribution of all models in Table 3 is shown in Figure 9 (A). Entries without accuracy/AUC values, or missing data were omitted. Also, for entries that report both values, accuracy was chosen to represent the model performance as it ais a common performance metric. Larger dataset sizes do not correspond to higher model performance, but some model types do appear to have been more prominently reported for some dataset sizes. Ensemble learning/RF, nearest neighbour algorithms, SVMs, and naïve Bayes models are more common where training datasets are under 1000 datapoints. Neural networks of all types see more use in datasets over 2000 datapoints.

When models with accuracy or AUC greater than or equal to 90% are considered, Ensemble learning/RF are more prominent among smaller datasets while neural networks, support vector machines and naïve Bayes are more common among larger datasets. Neural networks are the most represented algorithm type in this category, with three models scoring above 90%.

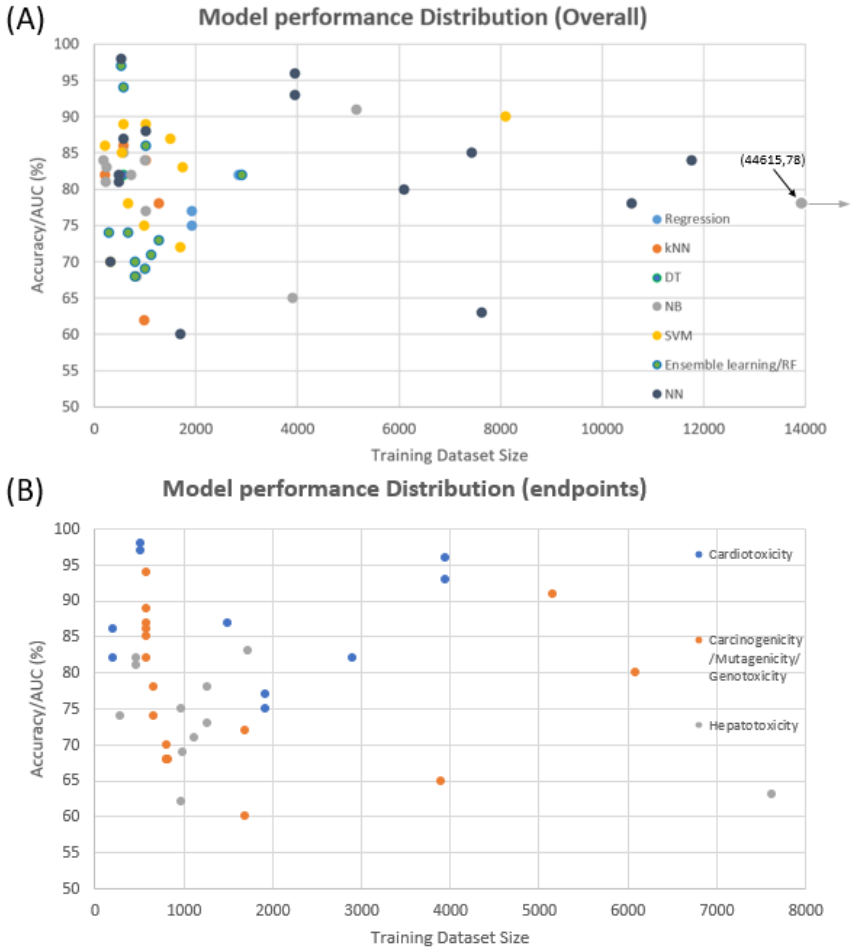


Figure 9. A graph showing model performance against training dataset size across (A) different model types (B) different endpoints.

Looking at the highest performing models for each of the toxicity endpoints most represented in Table 3, some suggestions can be provided for those new to machine learning model construction. The highest performing models for carcinogenicity, genotoxicity, and mutagenicity are random forests and support vector machines using MACCS keys and PubChem fingerprints as inputs and relatively small datasets.[180] High performing models on larger datasets use molecular descriptors and ECFPs in a naïve Bayes model.[149] For cardiotoxicity, the highest performing model is a deep neural network using fingerprints on a small dataset.[164] A support vector machine using MACCS fingerprints on a small dataset

provides the highest performing model in hepatotoxicity.[159] These are summarised in Figure 10 and serve as suggestions to be considered when constructing new models as the data type, distribution, and modelability also affect which models will perform best and how high model performance statistics will be.

It has to be acknowledged that it is hard to compare the algorithm performance over different toxicity endpoints due to the difference in complexity and data available. Hence, a subplot of Figure 9 (A) was generated for the three common toxicity endpoints, namely, hepatotoxicity, carcinogenicity/mutagenicity/genotoxicity, and cardiotoxicity. This subplot is shown in Figure 9 (B). There are more models predicting cardiotoxicity with an accuracy/AUC of above 90%, followed by carcinogenicity/mutagenicity/genotoxicity, and lastly hepatotoxicity. The lower overall performance of models predicting hepatotoxicity could possibly arise due to a lack of data used for building the models. It is also observed that most of the models have 2000 or less training data points.

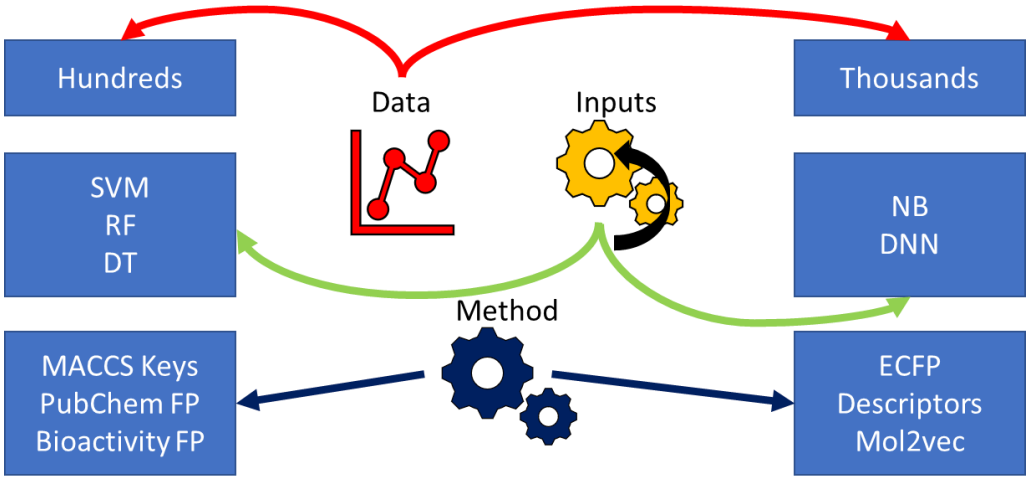


Figure 10. A summary of the highest performing methods and inputs found separated by dataset size.

To study model performance based on human vs animal and *in vivo* vs *in vitro* data, the studies in Table 3 were differentiated based on the data used where possible and plotted alongside each other (Appendices: Figures A1-A4). No clear trend or pattern was observed for the generated plots. This indicates that the model performance is dependent on the quality and quantity of the data used rather than the type of data (eg. animal vs. human). It is also possible that at this stage, there is still not enough data in the field to make a distinction between the model performance of different data types. Caution must also be taken when interpreting these results

because these studies use different datasets. The lack of high-quality *in vivo* data could also have contributed to this outcome.

In 2016, Huang et al built predictive models for 72 *in vivo* toxicity endpoints using *in vitro* data from the Tox21 10K library.[184] It was found that the *in vitro* assay data-based models performed better when predicting *in vivo* toxicity end points for humans than using *in vivo* data from animals to predict *in vivo* end points for humans.[184] Whilst more high-quality data for human toxicity and *in vivo* studies are required to better assess the model performance of these *in vitro* data-based models in the prediction of *in vivo* toxicity, this result demonstrates that *in vitro* data from cell-based assays offers a promising alternative to expensive and low throughput methods of obtaining *in vivo* data, and that extrapolation from these *in vitro* data to human *in vivo* end points is more reliable than extrapolation from animal to human *in vivo* data.[184]

In another study by Novotarskyi et al in 2016, the top model for the ToxCast EPA (Environmental Protection Agency) challenge was reported.[316] The aim of the challenge was to develop a model to predict the lowest effect level concentration (*in vivo* toxicity) based on *in vitro* measurements and calculated *in silico* descriptors.[316] A recent study by Xu et al in 2020 also investigated *in vivo* toxicity. In their work, predictive models for human organ toxicity based on *in vitro* bioactivity data and chemical structure were developed.[317] The models could be used to hazard screen large sets of chemicals for potential human toxicity, as well as to provide insights into toxicity mechanisms.[317]

In this chapter, *in vivo* and *in vitro* assay data are discussed with an expectation that the importance of *in vitro* data is likely to increase in future studies. A related subject is the extrapolation of *in vitro* assay data towards *in vivo* data (IVIVE) which could be important in the future of machine learning in predictive toxicology.[318]

The recent successes of machine learning in predictive toxicology have demonstrated that machine learning methods can generalise the data, as well as predict the potential toxicity of compounds accurately. While the results in Table 3 are generally of the single task classification method, multi-task classification/learning is also another method used in predictive toxicology. By learning tasks in parallel, multi-task learning has the potential to improve the generalization of the model, provided that sufficient data are available, longer training times and more complex architectures are practical, and that the distinct datasets are

sufficiently closely linked for the models to be related.[319] The assignment of more than one label to each instance might improve model performance by increasing model complexity. Several papers have explained multi-task classification, and this could be an alternative approach to take instead of varying the machine learning method used.[20, 319-321] For example, the work by Mayr et al in 2016 used multi-task learning and found that it enhances the model performance for 10 out of 12 assays.[18] In another study, Wu et al used multi-task learning on four different quantitative toxicity datasets.[11] Generally, it was observed that the multi-task models performed better than single task models when suitable data were available.[11] Hence, by using machine learning in predictive toxicology, the advantages of the various machine learning methods can be applied to the databases of drugs and drug-like compounds. This would lead to even more efficient and accurate predictions for drug toxicity.

However, there are some considerations when using machine learning to predict toxicity, and even for using machine learning in general. Firstly, care must be taken when processing the dataset for input into the model. This is because the results obtained by machine learning are highly dependent on the characteristics of the input data. For example, a dataset containing a significant majority of non-toxic compounds will likely result in a model that is skewed towards predicting non-toxicity. This would also affect the model's ability to predict toxicity on unseen data. This is part of the imbalanced data problem in predictive toxicology, as most of the available data are about toxic compounds, while the number of non-toxic compounds is significantly fewer. Other than the class distribution in the dataset, the size of the dataset also needs to be considered. Generally, machine learning methods perform better and can generalise better as the size of the dataset increases, provided that the quality of the data does not diminish. Another issue to take note of for machine learning concerns overfitting when training the model on the dataset. Overfitting of a model means that the model learns the training data too well; that is the model has memorised the training data. This affects the results when predicting new, unseen test data, in which case the model normally performs badly or more poorly than expected. In contrast, the results for the training data would score very well across most common performance metrics. During model training, overfitting can be identified as the region after the point at which the loss function reaches the minimum and is represented by an increase in the loss function after the minimum point. Another indicator of overfitting is when there is a significant difference between the training and test accuracy, or when the gap between these two metrics increases during model training.

Some methods to tackle overfitting include sampling techniques for imbalanced data, which were mentioned in an earlier section. In contrast, regularisation is commonly used during training to handle the overfitting problem. Regularisation aims to minimise the loss function, subject to a regularisation condition on the model parameter, where the regularisation parameter is represented by λ.[322-324] The first regularisation method is early stopping, which as its name suggests, is stopping the training of the model before the overfitting region.[325, 326] However, this can only be carried out if there is a clear identification of the overfitting region. Another two common regularisation methods are L_1 and L_2 regularisation. L_1 regularisation uses a penalty term which encourages the sum of the absolute values of the parameters to be small while L_2 regularisation encourages the sum of the squares of the parameters to be small.[327] More details about regularisation can be found in the literature. [322-331]

Therefore, overfitting is an issue of immense importance in the development of all computational models, and complex machine learning algorithms are often considered to have the most danger of overfitting. Modellers often use regularisation or external validation to limit the effects of overfitting, but these do not establish how applicable a model is to an incoming novel chemical. For this an applicability domain is appropriate, and these domains are common practice in the QSAR field and have been identified as a key element in *in silico* toxicology modelling.[332] The use of applicability domains does not yet seem to be commonplace in the development of machine learning algorithms.

With the vast quantity of data available, how should one then choose a machine learning model for their dataset? Perhaps, there is no best method that can generalise all datasets, but rather only the most suitable method. For data with a strong correlation between features, regression models, kNNs, NB models, or SVMs seem to be the most suitable due to their characteristics. DTs or RF can be considered for noisy data as they are generally more resistant to noise while being able to output results efficiently. To model complex problems which typical machine learning methods are unable to handle, deep learning (for example DNNs), as well as ensemble learning seem to be the most suitable machine learning method of choice due to their more complex model architecture. Images are handled by the specialised CNNs, while ANNs are a general machine learning method that can be used to model data. Therefore, understanding the data well is the first step to building a successful model in machine learning.

2.6 Future outlook and conclusion

Thus far, machine learning has been discussed, while the recent results of machine learning in predictive toxicology have been summarized and analyzed. In the future, it is expected that more models will be developed to predict toxicity, especially with technological advances which help lower computational costs and the continual development of new data sources. However, much needs to be done to address the main bottleneck facing machine learning in predictive toxicology, which is the quality and quantity of the data that is available to create datasets. While collaborations with pharmaceutical companies help mitigate part of this issue, as well as there being publicly accessible databases online, there are some gene or protein targets, or even toxicological endpoints which cannot be reliably predicted due to the lack of data. However, if a complete computational model, or more likely, a collection of computational models encompassing all of human toxicology, is to be built, these gaps in data need to be addressed. Moreover, more must be done to solve the issue of imbalanced data in predictive toxicology, an example of which is to collect, and disseminate, negative experimental results for the compounds.

Nevertheless, there have been many machine learning models that have high-performance metrics for predicting toxicity, which demonstrates the applicability of machine learning in predictive toxicology. However, even the best performing model type has its own set of limitations that has been covered and needs to be addressed and improved on, for machine learning in predictive toxicology to make further advances. While several common types of machine learning methods have been discussed in this thesis chapter, other machine learning methods are being developed and may become influential as their efficacy is established.

Mechanistic understanding is also key in the future of toxicology, hence, another topic on the rise in predictive toxicology deals with adverse outcome pathways (AOPs) and molecular initiating events (MIEs).[58, 59, 333-338] The lack of data on some AOPs and MIEs prove to be detrimental when trying to understand the prediction for these toxicological endpoints. Although there have been some improvements in this aspect such as the establishment of a publicly accessible AOP database (AOPWiki: https://aopwiki.org/), much still has to be done if the accurate understanding and prediction of toxicity is to become a reality.

Nevertheless, *in silico* methods have been increasingly employed in predictive toxicology, particularly during the screening process of new chemicals for safety decision making. The use of *in silico* methods such as machine learning to complement *in vitro* methods is the current status quo of the industry. While examples tend to be commercially sensitive and this data are rarely shared openly, there are some recent papers that contain some information about the use of *in silico* methods in the industry.[339-341] As the amount of available data increases, machine learning methods will likely become more popular due to their scalability. In particular, if more data can be generated or made publicly available, machine learning is also expected to perform better even if there are no improvements in the current algorithms. Perhaps, this indicates a possible direction of development for future *in silico* methods, where the focus will be on generating new data.

Despite their successful performance, the use of machine learning methods still remains a small part of the toxicological risk assessments in the industry and of regulators. It is acknowledged that such a cautious approach protects consumers of products from unnecessary risks, ensuring that the products can be used with ease of mind. It is important that the protocols for *in silico* methods are reliable and of good quality, which in turn leads to reproducible results. While *in silico* and *in vitro* approaches have recently experienced a period of development, there remains a need for their risk assessments to be as rigorous and understandable as those of traditional methods for them to be validated and accepted by regulatory bodies, and before they will be widely accepted as a replacement for animal testing. Currently, machine learning algorithms are combined with more traditional computational approaches such as read across and experimental *in vitro* studies as part of a weight of evidence approach. With more evidence to supplement these alternative non-animal approaches, regulators will have more confidence when making decisions. Recently, the new approach of next-generation risk assessments incorporates *in silico* methods as part of a weight of evidence approach which signifies that the industry is moving in the right direction.

Currently, machine learning algorithms are significantly useful, and this is highlighted by the recent successes of such methods in predictive toxicology. However, these methods still have much potential to be unlocked, which can only be done if the issues of insufficient high-quality data, regulatory acceptance, and model interpretability are resolved. Nevertheless, the future of machine learning applications in predictive toxicology is bright, and the hope is that *in silico*

methods, in particular machine learning algorithms, will be increasingly used in the industry and in academia to complement the use of *in vitro* methods.

Chapter 3: Knowledge transfer between different human targets in predictive toxicology using Tanimoto similarity and machine learning

3.1 Introduction

After a general introduction of machine learning in predictive toxicology was covered in Chapter 2, this chapter will have a focus on knowledge transfer between different human targets in predictive toxicology.

In recent times, the shift towards *in silico* methods for predictive toxicology has encouraged the use of several *in silico* techniques, one of which is machine learning. Machine learning is the generalisation of data, whereby patterns are drawn from a dataset. This generalisation would then be further applied to a secondary dataset. The degree of similarity between two distinct datasets affects the model performance and transferability of a machine learning or statistical model that has been trained on one of the dataset(s) and is then applied to the other.[342] If the degree of similarity between datasets is small, it would be expected that the model would not be able to predict the properties of the second dataset very well. Thus, by knowing the similarity between datasets, a decision can be made on when a model developed on one can be applied to the other.

Similarity methods have been used to construct applicability domains, of which either the Tanimoto similarity coefficient or the Pearson's coefficient can be used.[342] Alternatively, distance-based measures such as the Euclidean distance is used to determine the applicability domain. Approaches using Tanimoto similarity are classified under distance-based methods. For molecular fingerprints, Tanimoto similarity has been regarded as a good choice as a similarity metric.[343] These methods essentially determine if two molecules are close/similar to one another, in which case a threshold for the Tanimoto similarity can be specified where the pair of molecules are considered similar if their calculated Tanimoto similarity is at or above this threshold. Further information on these chemoinformatic classification methods and their applicability domain can be found in the literature.[342]

For quantitative structure-activity relationship (QSAR) models, the Organization for Economic Co-operation and Development (OECD) has defined the applicability domain as "a theoretical

region in chemical space encompassing both the model descriptors and modelled response which allows one to estimate the uncertainty in the prediction of a particular compound based on how similar it is to the training compounds employed in the model development".[344, 345] This means that one can determine if a new data point is within the applicability domain, which means that it can be reliably predicted from the current data. Other studies have worked on applicability domains, and these can be found in the literature.[346-350]

In predictive toxicology, the concept of applicability domain has traditionally been tied to QSAR models and much less important when considering other machine learning methods.[110] In this study, these methods are considered to be separate. In the literature, some studies have focused on quantifying the applicability domain, such as the molecule similarity-based method by Liu et al on melting point data.[351] The state of the art in the field is ensemble variance, which uses the average predicted results of an ensemble/group of QSAR models.[352] As discussed in the literature, this method is known to have high computational overhead which is especially prominent when using large datasets.[352]

Tanimoto similarity has been chosen to be used in a process to determine if a machine learning model trained on one dataset can be applied or transferred to other datasets, as compared to a focus on the applicability domain. This is much closer to transfer learning which involves a transfer of knowledge between domains.[353] In doing so, the goal is to find out about the relationship between different targets, as well as to determine if the data from other human targets can be used to supplement existing data for a particular target, thereby mitigating the lack of quality data for some of the targets in predictive toxicology.

In risk assessments for toxicology, the goal is to establish a safe dose/concentration while taking toxicological hazard and exposure into account while in predictive toxicology, the aim is to predict if the molecule/compound is toxic or non-toxic, followed by its exposure (for example via a pharmacokinetic model). Predictive toxicology has applications involving drug or drug candidates, consumer products, and foods.[1, 16, 24, 346-353] In the pharmaceutical industry, predictive toxicology is important as it can help mitigate the costs of drug withdrawals from the market, of which the cost was $2.59 billion in the USA in 2014.[1, 39] Such high costs due to drug withdrawals come about due to unexpected side effects during clinical trials, or even after they have been released into the market. Thus, the screening of potential compounds and their hazard prediction is important. Common endpoints in predictive toxicology include

hepatotoxicity, carcinogenicity, and cardiotoxicity.[1] These endpoints are hence the focus of several machine learning methods.[166, 168, 169, 178, 181, 354]

In this chapter of the thesis, the average similarity between datasets (S) was calculated with molecular fingerprints such as Morgan fingerprints or extended-connectivity fingerprints (ECFP) and with the use of Tanimoto similarity. This was inspired by the study by Mervin et al that uses a similar method for bioactivity data, though the authors adopted a distance-based approach.[355] Their study uses data from WOMBAT, as well as bioactivity data from PubChem and ChEMBL.[355] In comparison, toxicological data from ChEMBL and ToxCast which are well known toxicological databases,[61-63, 356, 357] has been used for the work in this chapter. Applying the method described in this chapter to different databases would likely result in different levels of performance. In the literature, another study has also employed Tanimoto similarity with a distance-based approach.[354] Their study uses melting point data while the work in this chapter uses data from toxicological databases. The differences in the type of data used would likely result in the methods of the studies not being completely applicable to the other data types. In short, this demonstrates that the method described in this chapter applies the similarity method to a different set of data/databases as compared to the literature.

The method described in this chapter is also able to estimate the validation/test accuracies of machine learning models given the training accuracies by using the average similarities between datasets (Figure 11). Although the use of Tanimoto similarity on molecular fingerprints is a common occurrence in the field, thus far, the similarity has not been used in conjunction with the results of machine learning models to understand model transferability, especially when considering datasets from different human targets in the field of predictive toxicology.

Also, in this chapter, this method is used to estimate model transferability or the relationship between two datasets for different human targets eg. (AChE vs. ADORA2A; where the method was able to estimate the test accuracy for a model trained on the AChE dataset that was used to predict on the ADORA2A dataset). This allows one to judge if a particular target has similar mechanism of actions as another target by using the similarity between datasets and could potentially identify previously unknown relationships between human targets. This could also potentially mean that some of the data for AChE can be used for ADORA2A.

In situations where the data is limited, machine learning methods are not expected to perform well. Hence, the hope is that the method is able to offer reasonable results when such situations are encountered as well as provide insights into the relationship between different human targets.

3.2 Methods

As this study is based on earlier work by the group,[358, 359] the datasets and part of the code (fingerprint function was adapted for use) used for this study were taken from https://github.com/teha2/chemical_toxicology/tree/master/NeuralNetworks-March2020.[358]

The datasets are comprised of data on human biological targets and were obtained from ChEMBL (version 23) and ToxCast.[356, 357] The SMILES were used to generate Morgan fingerprints using RDKit (version 03/2020) with varying fingerprint lengths.[360] All SMILES were kekulized and aromatic flags were switched off before fingerprint generation. Tanimoto similarities were calculated using RDKit (version 03/2020).[360] All code for the algorithm was run using Python 3 within Google Colab or in a Jupyter Notebook.[361] Lastly, all machine learning models were trained with a total of 100 epochs.

All code used for this study is available at https://github.com/Goodman-lab/Knowledge_transfer_with_Tanimoto_similarity.

3.3 Results and discussion

3.3.1 Workflow for the method

A hypothesis is made that similar datasets should have similar model performance since machine learning (ML) models generalise their dataset. Although one can calculate all similarity pairs for all molecules in two datasets and obtain a measure of average similarity using those values, this becomes difficult for larger datasets due to the number of calculations performed. As many targets are going to be analysed, a quick method to estimate the similarity between datasets is required.

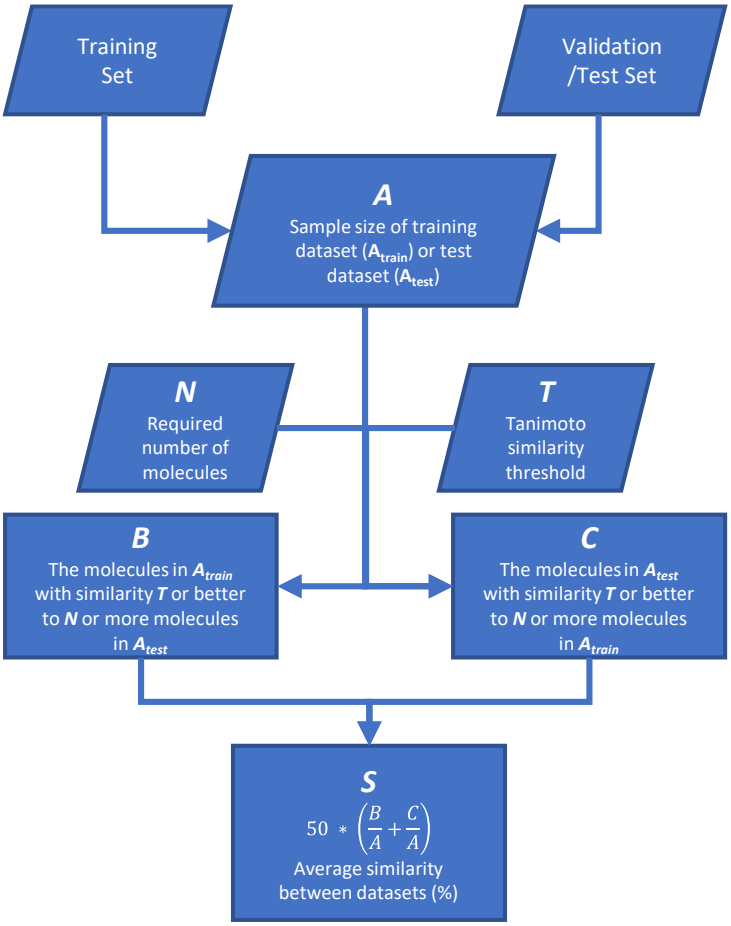


Figure 11: General workflow of the similarity method used in this study.

Figure 11 shows the general workflow for the method used in this study, where samples of the original dataset are taken and used to calculate the average similarity between datasets (S). The value of S depends on the key parameters/settings; the sample size of the datasets (A), the value of the Tanimoto similarity threshold (T) used, and the number of molecules before a molecule in one dataset is considered to be similar (N). This metric S is different from the average Tanimoto similarity of all fingerprint pairs between the two datasets as shown in Figure 11.

As the focus is on situations where data is limited, the choice of these parameters needs to account for this as well as maintain a balance between good results and the quantity of data in the datasets. If there is not enough data, this will result in a lack of similar molecules and thus poor results will be obtained. However, if there is too much data, there will be too many dissimilar molecules which would make the results worse while machine learning models will also generally outperform all other metrics in this situation. Next, the study will go on to show how the hypothesis is verified.

3.3.2 Optimal settings for similarity method

Since there are many datasets, data from five well-established targets (AChE, ADORA2A, AR, hERG, and SERT) were chosen to test and optimise the method. Each dataset was processed to fit the format required by the algorithm and fed into the algorithm. The results of the calculations are subsequently compared against the model test accuracies. The machine learning models are those that have been developed by Allen et. al for 79 human biological targets.[358]

The accuracy of the model for the training set (Q) can be calculated, as well as the value for S for the test set. The hypothesis is that the product of S and the accuracy of the training set (Q) will give a useful measure of how accurate the model will be when applied to the test set; SQ = P, where the product is the predicted test accuracy. This is useful because this relationship enables the calculation of the likely accuracy of a study using the test set without having to apply the model to the whole test set. This is particularly powerful if the test set is very large, such as all of the molecules in PubChem. To investigate this hypothesis, Figure 12 is presented which shows a plot of the absolute difference between P and the actual model test accuracy (DF) for a variety of values of S which were generated by varying N and T. The full results for Figure 12 are in Table A1 and Figure A5 in the Appendices.

By using this difference in P and the actual model test accuracy (DF), the accuracy of the method when calculating the properties of the test set can be determined. If the training and validation/test datasets are very similar, the value of DF would be small. On the other hand, if the value of DF is large, this would mean that the method predicts that the molecules in the training dataset are not representative of the test dataset *i.e.,* the ML model trained on the training dataset is not transferable to the validation/test dataset. Also, if there are a lack of molecules extracted by the method which occurs when the Tanimoto similarity threshold (T) used is too high, the predicted model test accuracy is affected because a "model" trained on a small number of molecules would be significantly different as compared to a ML model trained on a large number of molecules. As shown in Figure 12B, the average similarity between datasets (S) decreases as the Tanimoto similarity (T) increases.

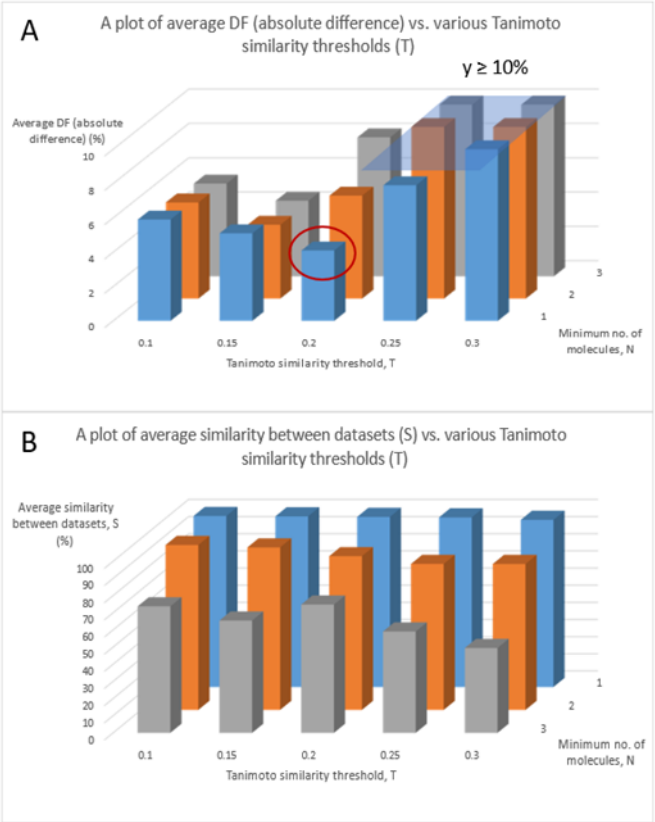


Figure 12: The absolute difference (DF) between the actual model test accuracy and the product P (SQ = P), where Q is the accuracy of the model for the training set, for several values of the Tanimoto similarity threshold (T) and the minimum number of molecules (N). (A) Plot of how the average DF varies with several values of N and T for the five targets; low values of DF show that the hypothesis is effective. N=1 and T=0.2 give the best results. (B) Plot showing how the average S decreases with increasing values of N and T. A high similarity is desirable. The actual model test accuracies were taken from the paper.[358]

In comparison, if DF is small, this means that the model trained on the training dataset is transferable to the test dataset. Thus, since the high values for the reported test accuracies indicates that the training and the test datasets used are similar, the optimal settings for the similarity method is tuned by minimising the value of DF. Also, by maximising the Tanimoto similarity coefficient (T), the method will use the most similar molecules to predict the test accuracy. A balance is needed to be struck between having enough data (low Tanimoto similarity) while having a high enough Tanimoto similarity to ensure the applicability of the machine learning (ML) model (similar molecules). In doing so, the aim is to ensure the method/process is able to measure the transferability of an ML model on a new dataset.

Figure 12 shows that the use of T = 0.2 and the use of N = 1 gives the smallest value of DF. This would result in efficient transferability of the ML model between the first dataset and the second dataset. Above this Tanimoto similarity threshold, the value of DF increases, indicating that increasing the Tanimoto similarity further results in the method predicting that the training dataset is not representative of the test dataset, and that the method cannot be representative of the actual ML model. Thus, the initial optimal setting means that the algorithm would calculate the similarity between two datasets, where the molecules in one dataset are considered to be similar if they have at least one molecule in the other dataset with a Tanimoto similarity of 0.2 and the same active state for both molecules in the similarity pairs.

As shown in Figure 12, the value of DF for a similarity threshold of 0.1 to 0.2 is somewhat similar, but this difference significantly increases when the similarity threshold is adjusted to 0.25. This indicates that for these datasets, using a threshold of 0.1 to 0.2 will be acceptable. A possible reason for this could be that the datasets are diverse, and thus using a high similarity threshold makes it difficult to find similar molecules between the datasets.

The method was tested by comparing the predicted values against calculated/published values for machine learning models. In doing so, the aim is to determine if the model results will be influenced more by the data or the model architecture. In addition, by knowing the similarity between datasets, one can determine if it is worthwhile to train a ML model.

Based on Table 4, using a value of T = 0.2 and N = 1 generally works for the five datasets tested for the optimisation process, with all targets producing values of DF of less than 10%. Also, the average similarity for each dataset was calculated using a sample size of 500. The average result of five runs was taken for each of the five targets and the results are summarised in Table 5.

The data in Table 5 demonstrates that the sample size of 500 is sufficient for further analysis as the standard deviation across all runs is low. Furthermore, the consistent sample size eliminates errors due to the size of the datasets being different across targets. For larger datasets, taking a sample of the dataset will reduce the computational cost. However, if the dataset is small, it would be advisable to use the entire dataset instead. The choice of when to sample the dataset, as well as how large a sample one should use, depends on the properties

and distribution within the dataset, which is not in the scope of this method/study as they are sampling techniques.[362]

Table 4: A comparison of the predicted test accuracies with the published accuracies of the machine learning models for selected human biological targets.

Target	Similarity, S (%)	Reported training accuracy (%)	Predicted test accuracy, P (%)	Reported test accuracy (%)	Absolute difference, DF (%)
AChE	91.2	91.4	83.4	84.7	1.3
ADORA2A	93.8	96.6	90.6	94.7	4.1
AR	87.5	93.2	81.6	90.1	8.6
hERG	90.9	93.9	85.4	77.3	8.1
SERT	93.5	98.6	92.2	96.0	3.8

The calculated similarity is between the training and test datasets for the [100,100] architecture. A similarity threshold of 0.20 and a fingerprint length of 10000 bits was used for the calculations in the algorithm. Also, although the paper uses ECFP4, the algorithm in this study uses Morgan fingerprints with a radius of 2, which is known to be comparable.[358] The similarities were calculated using a sample size of 500 for both the training and test sets. The reported/published accuracies were taken from the study.[358] Absolute difference (DF) measures the absolute difference between predicted (P) and model test accuracies.

Table 5: The average similarity between training and test datasets (S) for each of the targets with a total of five runs performed.

Target	Average similarity (%)	Standard deviation (%)
AChE	90.6	1.3
ADORA2A	93.0	0.9
AR	88.1	1.2
hERG	89.7	0.9
SERT	94.3	0.5

A similarity threshold of 0.20 and a fingerprint length of 10000 bits was used for the calculations in the algorithm. The paper uses ECFP4 but the algorithm in this study uses Morgan fingerprints with a radius of 2, which has been shown to be comparable.[358] The similarities were calculated using a sample size of 500 for both the training and test sets. Every dataset used was shuffled for each run.

3.4 Further applications of similarity method on more human targets

After the method has been optimised with the five targets, the method was also applied to more datasets/targets from the github/reference as well as testing the validity of this method on the remainder (74) of the 79 targets.[358] These results including that for the previous five targets are summarised in Table A2 in the Appendices which consist of a total of 79 targets. Based on Table A2, most of the results show good performance *i.e.* the value of DF is small (<10%). Overall, the accuracy rate for this method across all tested datasets is 98.7% where the outcome of the method is treated as successful when the value of DF is less than 10%. The average value of DF across all targets was also calculated to be (5.6 ± 2.1 %). This indicates that this method can reliably predict the test accuracy of a dataset given the training accuracy of a ML model and can be used to gauge the similarity or the model transferability between two datasets, even if the two datasets come from different targets.

In the work by Allen et al, the results for an architecture of [1000,1000] were also reported which were thus tested with the similarity method as well.[358] These results are summarised in Table A3 in the Appendices. Using the results in Table A3, the accuracy rate for the similarity method was determined to be 98.7% for the neural network architecture [1000,1000]. The average value of DF for the [1000,1000] architecture was also calculated to be (5.2 ± 2.1 %) which is comparable to the [100,100] architecture, although the average value is slightly lower. The consistent results of the method across both architectures have also highlighted the reproducibility and reliability of this method. This also means that one can trust the results predicted by the method 98.7% of the time.

Comparing the two architectures, the total number of targets having better performance (a smaller value of DF) as predicted by this similarity method for the [1000,1000] and the [100,100] architectures is 60 versus 19 respectively. The findings by the method in this study thus suggest that the [1000,1000] architecture is better than the [100,100] architecture. However, it is noted from Table A4 in the Appendices that the number of targets with the lower value of DF matching the better model architecture judged by the study is 42 of 79. This translates to a percentage of 53.2%, which is no better than random guessing. The differences between the results of the study by Allen et al and the method described in this chapter suggest that the difference between predicted and reported test accuracies (DF) as calculated by this method cannot be used to judge if a particular model architecture will outperform another

model architecture.[358] Such a conclusion is to be expected, as the information (features) in the datasets do not carry any information about the architecture and the model parameters. Hence, it is shown that the similarity method cannot make up for inadequacies or errors in building an optimised machine learning model.

The study by Allen et al has also reported model performance statistics for varying ECFP4 fingerprint lengths.[358] In order to test if the similarity method can be used to deduce the optimal fingerprint length required for each target, calculations were performed on the five targets reported in the study. The results of these calculations are summarised in Table 6.

As predicted, the similarity between datasets will increase as the fingerprint length decreases due to the higher proportion of "on" bits (bits with a value of 1), which increases the Tanimoto similarity for each similarity pair. A fingerprint length of 20000 bits has the highest value of DF while the value of DF experiences a decrease of 1.3% when decreasing the fingerprint length from 20000 to 1000 bits. This suggests that the fingerprint length does not significantly affect the information encoded in the molecular fingerprint, and that a smaller fingerprint seems to perform better. However, caution must be observed as changing the fingerprint length would affect the ability of the model to learn and generalise the data because the number of features is changed. Once again, this method does not provide any guidance on how a machine learning model should be built or optimised.

In order to demonstrate the potential of this method on other datasets, the data from a study by Zhang et al about drug-induced liver injury (DILI) was used.[181] The similarity method with the optimal settings was thus applied on these data, and the results are summarised in Table 7. It is observed from Table 7 that the values of DF are noticeably higher than those summarised in Table A2 and Table A3 in the Appendices where the similarity method was applied on human target data. Such a disparity is attributed to the use of a different type of fingerprint (substructure-based) by the work of Zhang et al,[181] while the similarity method chapter uses circular fingerprints. The difference in representations is thus the likely cause of the increased error. Another alternative explanation is that the data for DILI and the data for human targets are dissimilar, causing an increased error in the results because the method has only been shown to work for similar datasets thus far. Compared to similar datasets, the relationship between different datasets is more significant as it suggests the possibility of knowledge transfer using an existing model, or simply determining the relationship between two different targets.

Table 6: Results obtained by the similarity method for various ECFP4 fingerprint lengths for the five targets taken from the work by Allen et al.[358]

		1000 bits		5000 bits		10000 bits		20000 bits	
No.	Target	Similarity, S (%)	Absolute difference, DF (%)	Similarity, S (%)	Absolute difference, DF (%)	Similarity, S (%)	Absolute difference, DF (%)	Similarity, S (%)	Absolute difference, DF (%)
1	AChE	92.8	3.7	91.5	3.2	91.2	0.4	89.2	5.2
2	ADORA2A	94.1	5.0	93.1	5.7	93.8	5.6	93.6	4.5
3	AR	90.1	7.8	87.2	10.7	87.5	9.6	88.3	10.0
4	hERG	95.1	2.2	90.8	1.7	90.9	3.9	89.3	4.8
5	SERT	95.5	2.9	94.5	3.5	93.5	4.7	93.6	3.5
Average		93.5	4.3	91.4	5.0	91.4	4.8	90.8	5.6
Standard deviation		2.2	2.2	2.8	3.5	2.5	3.3	2.6	2.5

Table 7: A comparison of the predicted values obtained using the similarity calculations with the reported accuracies of the machine learning models from a study investigating DILI by Zhang et al.[181]

No.	Fingerprint used in study	Calculated similarity between training and test datasets, S (%)	Reported training accuracy (%)	Predicted test accuracy, P (%)	Reported test accuracy (%)	Absolute difference, DF (%)
1	FP4	79.0	67.2	53.1	65.7	12.6
2	MACCS	79.0	70.6	55.8	66.5	10.7
3	MACCSKeys	79.0	79.7	63.0	64.5	1.5

The predicted test accuracy (P) was calculated by multiplying the training accuracy by the similarity between the datasets (S). A similarity threshold (T) of 0.20 was used for the calculations in the algorithm and a fingerprint length of 10000 bits was used. A minimum number (N) of one similarity pair was set as the threshold during the calculation. The similarities were calculated using a sample size of 500 for both the training and test sets. The results shown in this Table were obtained by the study using various fingerprints and information gain (IG)(used as a feature selection technique) thresholds.[181] The model/published accuracies were taken from the study by Zhang et al.[181]

Thus far, the method has been tested in cases where the training and test datasets come from the same target *i.e.* the datasets are similar. In order to verify if the method will continue to work even when the datasets are dissimilar, each of the five well-established targets (AChE, ADORA2A, AR, hERG, and SERT) were tested against the data from all other remaining targets (78 per target). This results in a total of 390 (78 x 5) data points for each Tanimoto threshold tested, where each data point represents a combination of training and test datasets from different targets. All these results are shown in Figure 13 as the average and standard deviation of the absolute difference at each of the various Tanimoto similarity thresholds (T) used.

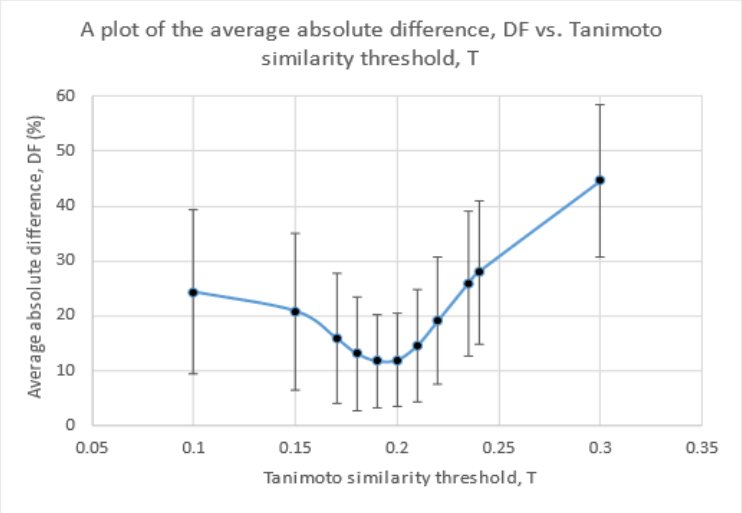


Figure 13: A plot of the average absolute difference (DF) across 3 runs (%) vs. various sample sizes for the training dataset of the AChE target vs. the test datasets of the five targets (AChE, ADORA2A, AR, hERG, and SERT). Every dataset used was shuffled for each run.

Based on Figure 13, both the average absolute difference (DF) across all 390 points and its standard deviation are close to the minimum at the threshold of 0.2, which also coincides with the initial optimal settings. Hence, this verifies that the threshold of 0.2 is the optimal choice for the algorithm, even when the training and test datasets come from different targets. Using

a slightly lower threshold (eg. 0.19) is not expected to have a significant impact on the results as observed from Figure 13. The sample size for both datasets was also varied in order to verify if this sample size of 500 is the optimal sample size for the algorithm.

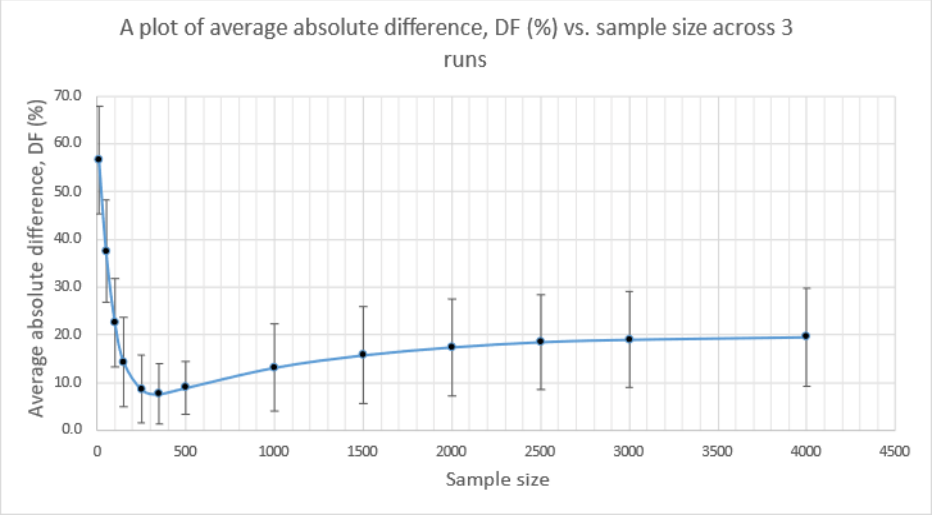


Figure 14: A plot of the average absolute difference (DF) (%) of 390 data points at various Tanimoto similarity thresholds (T) and N=1. From the plot, a sample size (A) of 500 is optimal and thus this was used for all subsequent analysis. The error bars represent the standard deviation.

The results of varying the sample size at the optimal threshold of 0.2 are summarised in Figure 14 which shows the effect of varying the sample size on the average value of DF using the AChE training dataset with the test datasets of the five well-established targets (AChE, ADORA2A, AR, hERG, and SERT). As expected, when the sample size is too small, the value of DF increases because there is not enough training data. However, as the sample size increases, the value of DF plateaus because the sizes of the original datasets are different, which results in some of the targets having a smaller training dataset than the sample size. Thus, this increases the value of DF for some of the targets and overall, this increase in DF will cancel out the advantages of using a larger sample size.

The plot shows that the sample size of 500 is the optimal choice for the algorithm since it has the smallest variance and is close to the minimum point. A two-sample paired t-test has also determined that the results for sample sizes 350 and 500 are the same. The test was performed at the 0.05 significance level where the null hypothesis is that the means for the two sample sizes are equal. Both the p-value for the one-tail and two-tail test which represents statistical

significance are calculated to be more than 0.05. Therefore, at this significance level, both one-tail and two-tail tests accept the null hypothesis that the means for the two sample sizes are equal.

Table 8: Results of the paired two sample t-test (for equal means) for the sample sizes 350 and 500. Data for the statistical test was taken from Table A14 in the Appendices.

SAMPLE SIZE	**350**	**500**
MEAN	7.623	8.893
VARIANCE	40.26	30.80
OBSERVATIONS	15	15
PEARSON CORRELATION	0.6018	
HYPOTHESIZED MEAN DIFFERENCE	0	
DF	14	
T STAT	-0.9185	
P(T<=T) ONE-TAIL	0.1869	
T CRITICAL ONE-TAIL	1.761	
P(T<=T) TWO-TAIL	0.3739	
T CRITICAL TWO-TAIL	2.145	

The AChE target was picked as it is one of the five well-established targets, and the trend shown in Figure 14 is expected to be the same even if another target is picked. This is because a sample size of 500 has been shown thus far to perform the best with the current algorithm and data. For the similarity method, an emphasis was placed on the speed at which the results are produced (ca. three days (sample size 500) vs. 25 days (all points in dataset) for calculating all similarity pairs). Thus, the current settings are sufficient for further usage and analysis, with the assumption that the sample size is representative of the entire dataset should a much larger dataset be used. However, for the toxicological datasets used in this study, the results from Figure 14 and Table indicate that one should use at least 350 data points, and that using more than 2000 data points doesn't contribute to a significant increase in model performance. As mentioned, and explained earlier in this study, we have chosen to use a sample size of 500. Regarding the "ideal" sample size, no clear guidelines have been published in the literature and it is largely dependent on the data used.

In Figure 14, the value of average DF roughly doubles from 500 data points to 2000 data points, showing that using too large a training set seems to result in overfitting, hence resulting in large errors. This could be due to the nature of this method, where the average similarity between datasets (S) and hence the predicted test accuracy (P) will increase as the size of the datasets increases. However, this increase in P will not correspond to an increase of similar magnitude for the actual model test accuracy. Thus, this would result in an increase in the value of DF as the size of the datasets increases to a large value.

The relationship between the predicted and actual model test accuracies were also investigated using data taken from Tables A2, and A5 to A9 in the Appendices. These results have been summarised and shown in Figure 15. The trendline has been plotted, together with the x=y line, which indicates the ideal case, in order to give a comparison. The average value of DF for the threshold of 0.2 is taken to be 12% which was determined from Figure 13 and is shown in Figure 15 as the blue region. In Figure 15, if the training and test datasets come from the same target, the datasets are regarded as similar and vice versa.

From Figure 15, it is observed that the gradient of the trendline for similar datasets is close to 0.54, which is slightly less than half of the gradient of the x = y line. This suggests that the method used in this study can relate the predicted test accuracy (P) to the reported test accuracy for similar datasets to a reasonable extent. However, the plot also shows that the method usually predicts a slightly lower accuracy than the results produced by the machine learning model. Since the predicted accuracy (P) is based on the similarity between the datasets (S), this implies that the method is slightly underestimating the similarity between datasets when the training and test datasets come from the same target. The small value of DF (<12%) hence demonstrates that this method can predict the validation/test accuracy well given a training dataset and the training accuracy, and assuming that the model has been optimised.

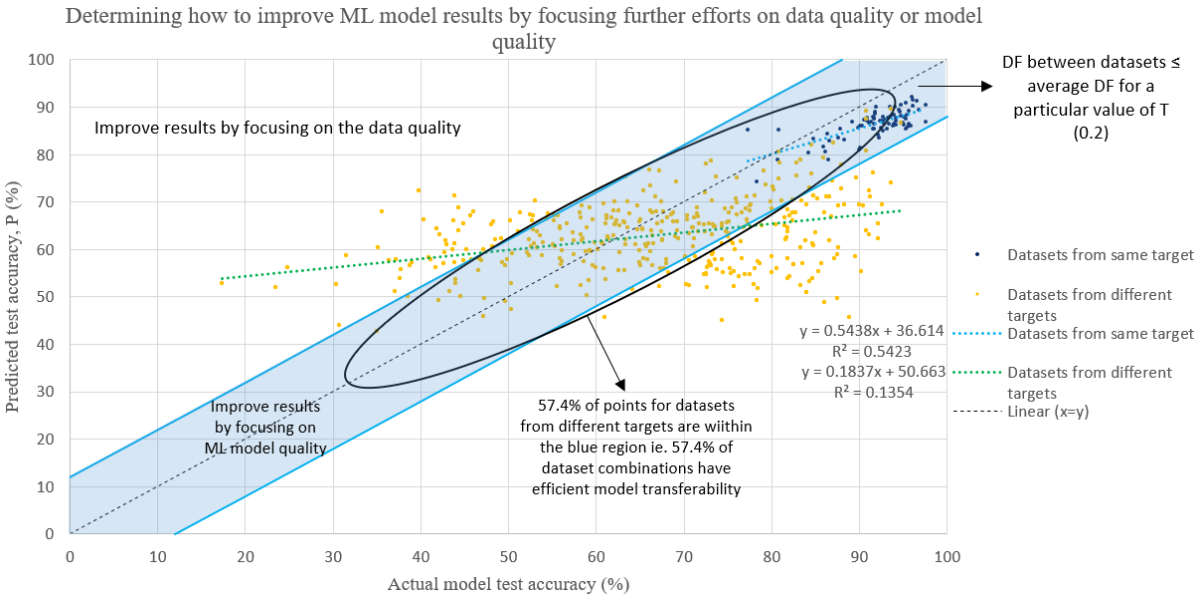


Figure 15: A plot of predicted test accuracy (P) (%) vs. actual model test accuracy (%) for datasets from various targets with T=0.2 and N=1. Each point on the plot represents a different combination of training and test datasets from the various targets. Training and test sets (datasets) either come from the same target or from different targets. Data for the plot were taken from Tables S2, and S5 to S9 in the Appendices. A sample size of 500 was used for all calculations. Additional plots for calculations using 0.1 and 0.3 similarity thresholds can be found in Figure S5 of the SI.

Comparing the results when the training and test datasets come from different targets to that when the training and test datasets come from the same target, it is observed that there is now significant variation between the predicted (P) and model test accuracy. This is due to the wide range of targets tested by the method which also demonstrates the viability of applying this method to different datasets. It was determined that 224 points (57.4%) are located within the blue region while the remaining 166 points (42.6%) are located outside the blue region. This means that out of all the 390 combinations of datasets tested, 224 combinations would be expected to work, that is the model trained on the data of the first target can be used to predict on the test data of the second target. Using an example to illustrate this point (using the data in Table A5 in the Appendices), the model for the AChE target can be applied on ADORA2A, ADRA2A, and ADRB1, but cannot be applied on AR and CHUK. This means that the former three targets are points that are located in the blue region while the latter two are points located outside of the blue region. Since the results show that the model trained on a target eg. AChE is able to predict on a different target eg. ADORA2A, this demonstrates that similar molecules

are expected to have similar behaviour which is the common assumption made when investigating structure-activity relationships. This does not mean that the targets have the same mechanism of action, but rather that the active site of the targets have some similarities which in turn results in similar molecules having similar activity for that target. This raises the possibility of using data from other targets to supplement the data for the current target, provided that the molecules are similar enough and their activities are the same. However, one must be cautious in doing so, especially if the toxicity endpoint is not well studied, as similar molecules do not necessarily have the same mechanism of action. The ideal approach to take would be to use a weight of evidence approach with other evidence such as results from *in vitro* methods. When the model trained on a target is unable to predict on another target eg. AChE is unable to predict on AR, this means that datasets and thus molecules for these two targets are dissimilar.

Also, the different regions as indicated on the plot in Figure 15 assist in determining if the obtained results are limited by the data or the ML model. For example, if the data is in the blue region, the training and test datasets are expected to have high similarity between themselves. Thus, the results obtained by an ML model are limited by the design and performance of the ML model since the quality of the data is good. In the other situation, when the data is expected to not be similar to each other (in the white regions), the results obtained by an ML model are limited by the data quality instead. Hence, Figure 15 demonstrates how one can use such knowledge to improve the results obtained by an ML model, even when the training and test datasets come from different targets.

In order to give a general idea of how to interpret the value of predicted test accuracy (P) without calculating the actual model test accuracy, Figure 16A was plotted. In Figure 16A, it is shown that as the predicted test accuracy increases, the proportion of points in the blue region (which represents efficient model transferability) increases. It is also noted that this probability of efficient model transferability increases significantly when the value of the predicted test accuracy is greater than 70%.

With this knowledge, it is possible to determine how much knowledge can be transferred when the training and test datasets come from different targets and could also be used to determine the relationship between targets. For example, the model trained on the AChE target produces an absolute difference (DF) of 4.5% with the data from the ADORA2A target, as well as a

calculated similarity (S) of 65.3%. This would mean that the knowledge of 65.3% of the AChE data can be transferred to the ADORA2A dataset. With the calculated similarity (S), one could infer if the two targets have a similar mechanism of action, since similar structures would tend to have the same activity *i.e.* the structure-activity relationship.

Also, in Figure 16A, the difference between the predicted and actual ML model test accuracy (DF) is plotted against the average similarity S (workflow in Figure 11) for training and test datasets from different targets. Using Figure 16A, it is possible to estimate if a model will be transferable on another dataset. Similar to the predicted test accuracy (P), as the average similarity between datasets (S) increases, the proportion of points in the blue region (Figure 15) which is related to efficient model transferability increases. The plot shows that that having an average similarity between datasets (S) of 70% or higher is expected to result in good model performance.

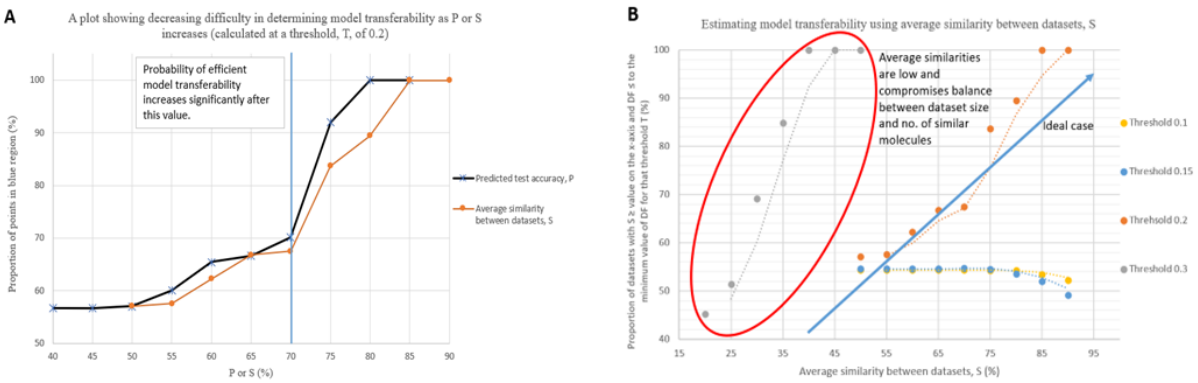


Figure 16: (A) A plot showing the change in the proportion of data points in the blue region of Figure 15 as the average similarity between datasets (S) or the predicted test accuracy (P) increases. The plot shows that having values of S or P at 70% or higher is expected to result in good model performance. Data for the plot was taken from Tables A2, and A5 to A9 of the Appendices. A sample size of 500 was used for all calculations. (B) A plot of proportion of datasets with S ≥ value on the x-axis and DF ≤ to the minimum value of DF for that threshold (%) (From Figure 3) vs. the average similarity between datasets (S) (%). Each point on the plot represents the proportion out of a total of 390 data points eg. At S = 55, about 57.5% out of 390 dataset pairs have a average similarity (S) more than or equal to 55% and with absolute difference (DF) less than or equal to 12% for the Tanimoto similarity threshold of 0.2. High average similarities between datasets are likely to result in good model performance provided the correct values of T and N are chosen.

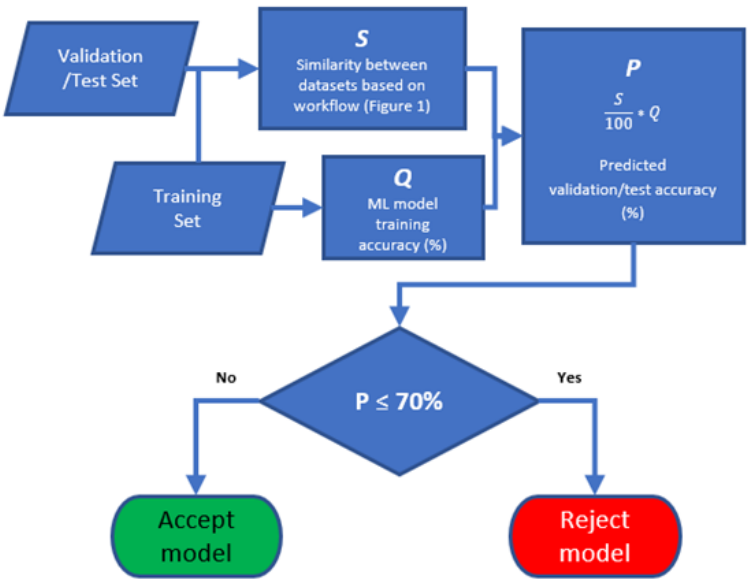


Figure 17: General workflow of applying the method in this study to two datasets.

The general workflow for applying the method in this study is shown in Figure 17. Worked examples for Figure 17 can be found in Figure A6 and Figure A7 in the Appendices. The workflow shows how one can easily determine if a model is transferable to another dataset.

However, with this similarity method, there are issues that one must account for. These include the presence of activity cliffs within the datasets, where an activity cliff is when a compound has a significantly different activity than the other compounds similar to it. This would cause issues with the machine learning model as it makes the data more difficult to generalise, though the model might treat these points as outliers.

Also, the use of similarity cannot determine if the dataset is challenging for a machine learning model at a glance. Once again, this is related to the ability of the machine learning model to generalise the data, of which no information about this is encoded in the dataset or features. Thus, this will also affect the accuracy of the similarity method. However, if the predicted test accuracy (P) is significantly different from the actual model test accuracy, it suggests that either the dataset is challenging to model, or that the two datasets have different feature spaces. If the predicted test accuracy is higher than the actual model test accuracy, this implies that the model is underperforming and is probably not optimised, or that the current model is overfitting to the training data. On the other hand, if the predicted test accuracy (P) is lower than the actual model

test accuracy, this implies that the difference in the observed accuracy is due to other factors other than the similarity between molecules.

Thus far, the method has been shown to be useful in determining the validation/test accuracy of a second dataset given the training accuracy. This would assist in determining if the obtained results are due to the data or the model. The similarity method could thus be used to determine the relationship between two datasets, where the calculated value of S can be taken as the measure of the relationship. As the similarity method can be used to gauge how much information can be transferred between two datasets, if the value of S is high, the machine learning model is likely to produce a good result when predicting on the test set (Figure 16). Naturally, this does not offer information about what the machine is learning, as this is once again not encoded in the information of the molecular fingerprint. Caution must be exercised as a high similarity method value and a good machine learning result does not necessarily mean that the machine learning model is predicting the "right" results.

In comparison to the study by Mervin et al which shows that 0.3 is the best Tanimoto similarity threshold to use for their datasets,[355] the results of this study show that 0.2 is the best Tanimoto similarity to use for the datasets in this study. This shows that it is perhaps better to try out both 0.2 and 0.3 similarity thresholds for a particular dataset, as they might offer better or different results. It is likely that the ideal Tanimoto similarity threshold depends on the data used. Also, the difference arises because a higher cut-off restricts the number of molecules in the training set. On the other hand, while the lower cut-off increases the number of dissimilar molecules, this disadvantage is outweighed by the increase in the total number of molecules.

This difference also suggests that toxicological data is very complex and using a single Tanimoto threshold for all datasets is underestimating the problem. If the dataset is diverse, and thus the similarity between molecules in the datasets are expected to be low, it is better to use a lower Tanimoto similarity threshold as this allows more data to be retained while keeping the similarity high. Conversely, if the dataset has a large number of similar molecules, this allows one to use a higher similarity threshold without compromising on the quantity of data used. Essentially, this is a situation of finding the right balance between the quantity of data and the similarity threshold used.

3.5 Further applications of similarity method on data curation

The method could also be used to determine which molecules in a dataset are likely to be the sources of error when the model makes predictions on them, simply because the model has not been trained on similar data. The solution to this is to either omit these test molecules or obtain more sources of data which have similar molecules to these. If such data is not available, this highlights a need for more data regarding these molecules to be obtained in order to make better predictions for the particular target. By identifying which data points the machine learning model is unable to generalise upon, the model can be improved as well as offer insights into where future data collection needs to be focused upon. Hence, the similarity method can be used to analyse and improve the results of the machine learning model.

Thus, this was investigated by choosing a well-established target, AChE, and testing it on the five targets chosen during the optimisation process. The results of a total of 10 runs have been summarised in Table 9. For each target, the data for both the raw and curated data, where the curated data refers to the raw datasets which have dissimilar molecules removed. The dissimilar molecules refer to those that do not meet the criteria of having a Tanimoto similarity above the specified threshold (0.2) and having the same active state. Both the training and test sets were curated, with all other model parameters remaining the same throughout the calculation. The full results can be found in Table A10 in the Appendices.

The results from Table 9 show that curating the dataset according to the method increases the test accuracy of the model by 22.5%, even if the two targets/datasets are distinctly different. This means that by curating the dataset, one is able to identify the data points which need further data before they can be modelled, as well as improving the models for those data points that can be modelled. This also potentially allows one to use data from other targets to increase the training data for a machine learning model, which alleviates the issue of insufficient data for some targets in the field of predictive toxicology. However, it is noted that curating the data to fit the test set means that it is more likely that the model will fail to reliably predict results on molecules that are not present in the training data. One should also consider if the predictions on molecules that are not present in the training data, in particular when the molecules are dissimilar, can be trusted given that the machine learning model has not seen enough of these molecules to generalise a pattern.

Table 9: A comparison of the test accuracies for the model of the AChE target with various targets for both the raw and curated data.

Test target	Raw data test accuracy (%)	Curated data test accuracy (%)	Difference in accuracies (%)
AChE	86.4	87.9	1.4
ADORA2A	56.4	87.8	31.5
AR	73.0	86.7	13.6
hERG	55.3	87.2	31.9
SERT	54.3	88.5	34.2
		Average	22.5

Curated data refers to the raw data in both training and test datasets without the molecules that have no other similar molecules by following the workflow in Figure 11. The difference in accuracies (%) takes the curated data as a reference and is calculated by subtracting the test accuracies of the raw data from the curated data. The results shown here are an average of 10 runs using the optimal settings. All the data was used unless specified otherwise. A Tanimoto similarity threshold of 0.2 was used.

3.6 Further applications of similarity method on toxicological databases

Moving on, this method has been applied to data from the PubChem and ChEMBL databases and the results are shown in Figure 18 and 19.[356, 363] The datasets from PubChem were compared with the datasets from the human targets using the workflow in Figure 11.

The plot in Figure 18 shows that as the number of molecules in the PubChem dataset increases, the average proportion of similar molecules does not change significantly. With this result, the prediction is that even if the entire PubChem database is used, this value will not change. Such a result indicates that if the models by Allen et al were used,[358] the behaviour of about 13% out of 100 million molecules and their toxicological effects can be predicted. This is because similar molecules are known to have similar behaviour if the structure-activity relationship holds. A similar result was observed for the ChEMBL database as well (Figure 19). If the models by Allen et al were used,[358] the behaviour of about 18% out of 2 million molecules and their toxicological effects can be predicted. The results from these two plots indicate the high toxicological potential of these popular toxicological databases.

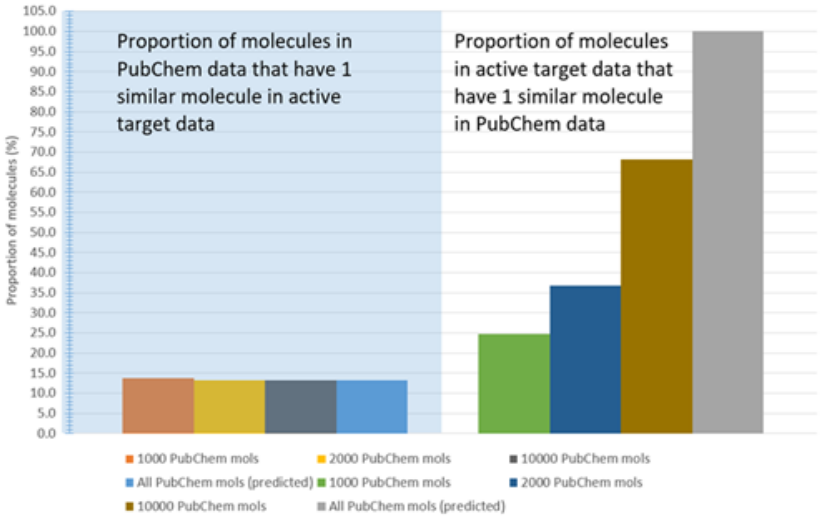


Figure 18: A plot showing the average proportion of molecules in one dataset that have one similar molecule in the other dataset. The average proportion was determined by taking the average across all 79 targets for a total of 3 runs. The molecules in the PubChem datasets were randomly extracted from the PubChem database with metals removed, salts stripped, and duplicates removed. This was repeatedly performed until the specified number of molecules was reached. A Tanimoto similarity threshold of 0.3 was used. (Full results in Figure A11 to Figure A13 in the Appendices)

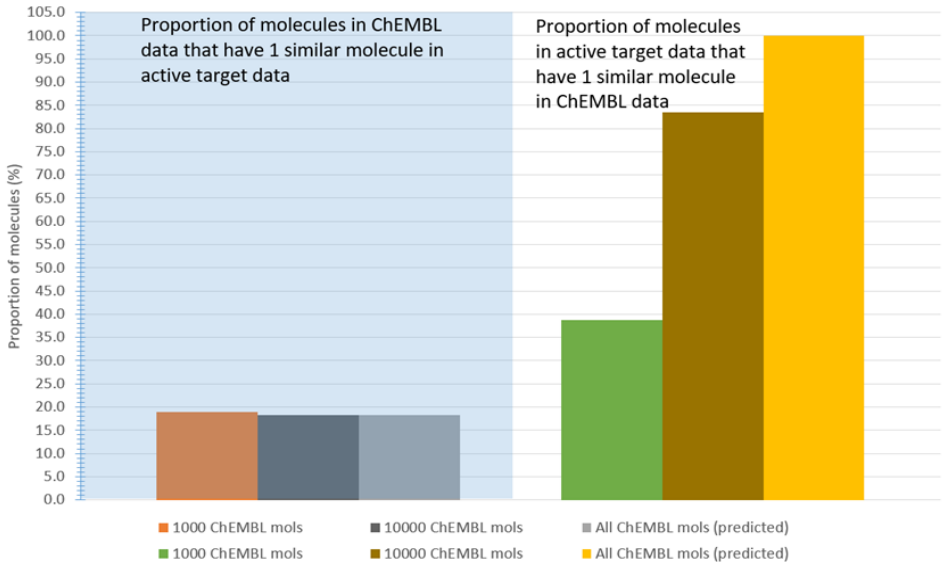


Figure 19: A plot showing the average proportion of molecules in one dataset that have one similar molecule in the other dataset. The average proportion was determined by taking the average across all 79 targets for a total of 3 runs. The molecules in the ChEMBL datasets were randomly extracted from the ChEMBL database with metals removed, salts stripped, and duplicates removed. This was repeatedly performed until the specified number of molecules

was reached. A Tanimoto similarity threshold of 0.3 was used. (Full results in Figure A14 and Figure A15 in the Appendices).

3.7 Conclusion

In conclusion, a method that uses Tanimoto similarity to determine the transferability of a model between two datasets was developed and applied to toxicological data for different human targets. This method has been shown to allow one to obtain a broad idea of the chemical space of a dataset, as well as the relationship between two datasets. This study has also demonstrated that with this method, when the average similarity between datasets (S) or the predicted test accuracy (P) is 70% or higher, it is expected that model performance will be good *i.e.* the model will be transferable to another dataset. This study has also shown that the proportion of similar molecules can be used for popular toxicological databases, where about 13 million molecules (PubChem: 13%, ChEMBL: 18%) can be analysed using the Allen et al models for a series of important toxicological effects.[358] Finally, the hope is that that this method/workflow can be used for other datasets in predictive toxicology.

Chapter 4: Investigating developmental and reproductive toxicity with the use of automated machine learning and transfer learning

4.1 Introduction

In Chapter 3, the knowledge transfer between different human targets was investigated. In this chapter, developmental and reproductive toxicity will be investigated, and transfer learning will be carried out to determine if any insights can be gained.

Developmental and reproductive toxicity (DART) has been the focus of recent research due to its importance in human health hazard and risk assessment. It is a complex endpoint consisting of both developmental and reproductive aspects. For the developmental aspect, this includes fetal development, teratogenicity, transfer across the placenta barrier, and prenatal developmental toxicity.[364-367] For the reproductive aspect, this includes gametogenesis, release of gametes, formation of the zygote, parturition of offspring, and lactogenesis.[367, 368] As compared to reproductive toxicity, developmental toxicity is also known to be more difficult to assess. This is because the developmental process is especially sensitive to genetic errors and environmental disruptions, windows of vulnerability being created by the timing of the processes, and the maternal effects having an impact on all stages of fetal development.[369] There is also concern about the inter-species difference and the reproducibility of the test results.[369] It is thus known that modelling the endpoint of DART is a complex and tricky task.

Several guidelines have been developed by the Organisation for Economic Co-operation and Development for the testing of DART. This includes OECD 414-416, 421, 422, 426, and 443.[370] However, these guidelines involve testing on many animals while also being costly and which requires a long testing time.[371, 372] For example, OECD 414 is used for the investigation of prenatal developmental toxicity and requires at least four groups of 20 female animals per compound tested.[370] If a traditional two-generation test is used, this requires up to 3200 animals and is among the most costly tests.[369] In 2013, the European Union imposed a ban on the sale and importation of cosmetics that were tested on animals or contain animal-tested ingredients.[372, 373] Consumers are also increasingly favouring products that have been developed with alternative non-animal methods.[374] As a result, there is an ongoing shift from animal testing to alternative non-animal testing methods in predictive toxicology.

It is important to develop non-animal approaches such as *in silico* methods to assess the potential DART toxicities of chemicals. These *in silico* methods can be used in a weight of evidence approach under the recent next generation risk assessment (NGRA) framework.[375, 376] There are multiple studies in the literature which have developed *in silico* methods for DART. The study by Wu et al in 2013 remains one of the popular choices for DART which covers a framework to identify structural alerts using an empirically based decision tree.[377] This allows one to determine whether or not a chemical has structural features that are consistent with chemical structures known to have toxicity for DART endpoints.[377] Teratogenicity is also part of DART and has been studied by Challa et al in 2020.[364] They have found that drug structure is a good predictor of teratogenicity and have developed a workflow to identify teratogenic moieties.[364] In 2021, Feng et al developed ensemble learning models for reproductive toxicity.[368] Their findings indicate that ensemble learning can play a beneficial role in the prediction of reproductive toxicity and can significantly improve the ability to predict such toxic compounds.[368] More recently, Ciallella et al has investigated prenatal developmental toxicity using chemical structures and biological data.[365] Their work has automatically identified bioassay data relevant to prenatal developmental toxicity from public databases and created a new strategy for predicting untested chemicals early in the discovery and development procedure or existing chemicals with limited safety data.[365] Despite all these studies, the investigation and modelling of DART using *in silico* methods with publicly available information remains a challenge.

In this study, the construction of the largest known database with 3245 compounds with valid QSAR ready SMILES (1662 positives, 1583 negatives) for the general prediction of DART is reported. The database construction process is explained in detail under the Methods section and is available as an Excel workbook, which contains a database with salts and metals included for reference purposes, and another database without salts and metals which can be used for modelling. These data were taken from a total of 12 sources covering a range of endpoints on DART. This database was subsequently used to train machine learning (ML) models using an automated learning (AutoML) process and the results are reported in this study. The misclassification of the models was also investigated in order to determine where the current models fall short, and where more data needs to be gathered.

Finally, the last part of this chapter will focus on transfer learning. Transfer learning is the development of a high-performing model for a target domain trained from a related source

domain.[378] This is usually carried out when the data for the target is limited and thus a model cannot be efficiently trained on the data. It is known that when there is a difference in data distribution between the training and test data, the results of a predictive model can be degraded.[378,379] Therefore, by using this property of transfer learning, the relationship between different toxicity endpoints can be investigated, or if certain human targets are involved with the toxicity endpoint. For example, the androgen receptor is known to be linked to reproductive toxicity.[380] Thus, a model trained on reproductive toxicity data would likely be able to predict well on androgen receptor activity data, especially if the reproductive data contains molecules that are related to androgen binding.

Naturally, transfer learning also has its limitations. While a good performance by the machine learning model when predicting on the test data indicates that the two data sources are related, poor performance results could either mean that the two data sources are unrelated, or that the model was trained on insufficient data to be able to pick out patterns effectively on the test data. Thus, one must be cautious when drawing conclusions about negative relationships between the data sources.

In this study, the similarity method in Chapter 3 was also used and the results obtained with transfer learning between developmental toxicity and reproductive toxicity in this chapter match the conclusion made in Chapter 3 about the similarity method, which is that when the average similarity between datasets (S) or the predicted test accuracy (P) is 70% or higher, it is expected that model performance will be good *i.e.* the model will be transferable to another dataset. Additionally, transfer learning between the models for developmental toxicity and reproductive toxicity with the human target data in Chapter 3 was carried out. In doing so, it is hoped that more insights can be gained into the relationships or mechanisms of action of both developmental toxicity and reproductive toxicity which are still not fully understood.

4.2 Methods

The raw data for the DART database was taken from the sources detailed in Table 10. The entry for each compound in the database was further supplemented with additional details such as the chemical name, InChI Key, and QSAR ready SMILES using the EPA CompTox Chemistry Dashboard: https://comptox.epa.gov/dashboard/. Next, entries with missing or unclear data were removed. If the compound has enantiomers, the toxicity values were merged

with the toxic values having priority over the non-toxic values for each source. For each source, the compound's toxicity value is shown as active (1), inactive (0) or not present (#N/A). Entries were determined to be unique (no duplicates) based on their name, InChI Key, and SMILES.

To prepare the dataset for modelling, salts, and inorganics (metals) were removed from the SMILES of the compounds. These SMILES will thus be called QSAR ready SMILES. If multiple compounds have the same QSAR Ready SMILES, the toxicity values for the sources are merged into a single entry, with the toxicant/toxic value having a higher priority should the compounds be tested for the same source. Entries without QSAR Ready SMILES were removed. This database without salts and metals is henceforth referred to as the DART database. The DART database contains a total of 3245 compounds (1662 positives, 1583 negatives). In the modelling process, only the QSAR Ready SMILES of compounds will be used, together with their overall toxicity value. The list of data sources in the DART database for the dataset without salts and metals is shown in Table 10.

Table 10: List of data sources for the DART database without salts and metals

No.	Data source	*in vitro*/*in vivo*	Developmental/Reproductive toxicity	No. of positives	No. of negatives	Total no. of compounds	Ref.
1	Collection of chemicals from the literature	*in vivo*	Developmental	38	4	42	381-422
2	Framework for identifying chemicals with structural features associated with the potential to act as developmental or reproductive toxicants	Mixed (multiple sources)	Developmental/Reproductive	583	42	625	377
3	Identifying reference chemicals for thyroid bioactivity screening	*in vivo*	Developmental	27	4	31	423
4	Development of a curated Hershberger database	*in vivo*	Reproductive	25	19	44	424
5	Development and validation of a computational model	*in vitro*	Reproductive	24	35	59	425

	for androgen receptor activity						
6	Identification of candidate reference chemicals for *in vitro* steroidogenesis assays	*in vitro*	Reproductive	22	48	70	[426]
7	Profiling the ToxCast library with a pluripotent human (H9) stem cell line-based biomarker assay for developmental toxicity	*in vitro*	Developmental	194	807	1001	[427]
8	A novel human stem cell-based biomarker assay for *in vitro* assessment of developmental toxicity	*in vitro*	Developmental	23	15	38	[428]
9	EPA ICE data (provided by Dr Katarzyna R. Przybylak)	*in vitro*	Developmental	20	10	30	-
10	Predicting the reproductive toxicity of chemicals using ensemble learning methods and molecular fingerprints	*in vivo*	Reproductive	744	793	1537	[368]
11	Machine learning on drug-specific data to predict small molecule teratogenicity	Mixed (multiple sources)	Developmental	416	153	569	[364]
12	Predicting prenatal developmental toxicity based on the combination of chemical structures and biological data	Mixed (multiple sources)	Developmental	588	516	1104	[365]

Some compounds in the literature were selected because they were interesting. These compounds are grouped together as a collection of literature and shown as entry 1 in Table 10. All compounds are tested for developmental toxicity *in vivo*. Some of these compounds are also not found in other data sources. Entry 2 in Table 10 is the work by Wu et al which covers a framework to identify structural alerts using an empirically based decision tree.[377] This allows one to determine whether or not a chemical has structural features that are consistent with chemical structures known to have toxicity for DART endpoints.[377] Entry 3 is a study by Wegner et al where reference chemicals were selected based on thyroid bioactivity in 'Tier 1' screening assays used by the US EPA's Endocrine Disruptor Screening Program.[423] Entry 4 covers the work done by Browne et al where a systematic literature review was conducted to identify Hershberger bioassays for chemicals including those used to validate the OECD/US EPA guideline assay, US EPA's chemicals screened for endocrine activity, and the library of chemicals run in US EPA's ToxCast *in vitro* assays.[424] The work by Kleinstreuer et al. is listed as entry 5 and is about the integration of 11 high-throughput *in vitro* screening ToxCast/Tox21 *in vitro* assays into a computational network model to distinguish true androgen receptor pathway activity from technology-specific assay interference.[425] Entry 6 describes an approach by Pinto et al for identifying *in vitro* candidate reference chemicals that affect the production of androgens and estrogens in models of steroidogenesis from a review of the ToxCast high-throughput H295R steroidogenesis assay and gonad-derived *in vitro* assays used in methods validation and published in the scientific literature.[426]

The work by Zurlinden et al is listed as entry 7 and is about the Stemina devTOX quickPredict platform which is a human pluripotent stem cell-based assay that predicts the developmental toxicity potential based on changes in cellular metabolism following chemical exposure.[427] Entry 8 is about the development of a new assay by Jamalpoor et al, ReproTracker, which is a state-of-the-art *in vitro* method that can identify the teratogenicity potential of new pharmaceuticals and chemicals and signify the outcome of in vivo test systems.[428] Data for this entry was taken from two sources, namely Table 1 of the work by Jamalpoor et al using the ReproTracker assay, where the recorded value in the DART database is taken to be positive according to the values of teratogenicity in Table 1,[428] and the ReproTracker website (https://toxys.com/reprotracker/) where the toxicity value of a compound is positive if the picture shows a red cross and negative if the picture shows a green tick. These can be found under the "Validation" tab on the ReproTracker website. The EPA ICE data on in vitro developmental toxicity which was provided by Dr Katarzyna R. Przybylak is listed in Table 10

as entry 9. Entry 10 covers the work done by Feng et al who developed ensemble learning models for reproductive toxicity and whose findings indicate that ensemble learning can play a beneficial role in the prediction of reproductive toxicity as well as significantly improve the ability to predict such toxic compounds.[368] Entry 11 is about the work by Challa et al who have found that drug structure is a good predictor of teratogenicity and have developed a workflow to identify teratogenic moieties.[364] Lastly, entry 12 describes the work done by Ciallella et al on prenatal developmental toxicity using chemical structures and biological data.[365] Their work has automatically identified bioassay data relevant to prenatal developmental toxicity from public databases and created a new strategy for predicting untested chemicals early in the discovery and development procedure or existing chemicals with limited safety data.[365]

Finally, Table 10 contains a total of four entries with *in vivo* data, five with *in vitro* data and three with mixed data. There are 1621 chemicals for *in vivo* data (878 positives, 743 negatives) and 1064 chemicals for *in vitro* data (439 positives, 625 negatives). In Table 10, seven entries cover developmental toxicity, four cover reproductive toxicity and one covers both developmental and reproductive toxicity. The main endpoints that the database covers for developmental toxicity include thyroid bioactivity, teratogenicity, and prenatal developmental toxicity. For reproductive toxicity, the database covers androgen-responsive endpoints, steroidogenesis, sperm reduction, gonadal dysgenesis, abnormal ovulation, teratogenicity, infertility, and delayed growth. Developmental toxicity has 2202 chemicals (1206 positives, 996 negatives) while reproductive toxicity has 1606 chemicals (848 positives, 758 negatives). All entries in Table 10 predicting for both toxicity endpoints and/or containing mixed *in vivo*/*in vitro* data were excluded from the numbers reported in this paragraph.

The overall toxicity value of each compound in the database was determined based on the threshold used. Two thresholds have been defined in the Excel workbook containing the database. The first being that the overall toxicity value is positive if any of the recorded sources are positive. The second threshold treats the overall toxicity value as positive if the number of positives for the recorded sources are greater than the number of negatives for the recorded sources eg. if a compound was tested in seven sources of data, with three positives and four negatives, the compound would be treated as being non-toxic. In the consideration of these thresholds, "#N/A" values are not considered. Unless otherwise specified, the threshold used for the overall toxicity of each compound is when the overall toxicity value is positive (1) if any of the recorded sources are positive (1).

An external test set was processed similarly to the DART database. Compounds in this external test set that were also found in the DART database were removed from this test set. This ensures that all compounds in the external test set are not present in the DART database as well as the training set for the DART models. This dataset was obtained from Hewitt et al and contains 290 compounds for developmental toxicity.[429] After processing, the external test dataset contains a total of 68 chemicals (41 positives, 27 negatives).

SMILES for the compounds were converted into Morgan fingerprints with 2048 bits and radius 2. This process was carried out using RDKit (version 2022.03.1). Unless otherwise stated, a total of five runs were performed for the DART database with 24 models being trained each run and the results are reported as an average of these 5 runs. This means that a total of 120 models are trained and evaluated for a dataset. During each run, 20% of the dataset was randomly split to be the test set. The remaining 80% of the dataset was then used for model input in the AutoML process, with 5-fold cross validation being used. For each fold, the training and validation split is 80% and 20% respectively. The data was also shuffled before each run.

To evaluate the model performance, several common metrics were used. These consist of accuracy (ACC), sensitivity (SE), specificity (SP), and the Matthews correlation coefficient (MCC) which are calculated using the values for true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). The formulas for these metrics are shown in equations (9) to (12).

$$ACC = \frac{TN+TP}{TN+FP+TP+FN} \tag{9}$$

$$SE = \frac{TP}{TP+FN} \tag{10}$$

$$SP = \frac{TN}{TN+FP} \tag{11}$$

$$MCC = \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{12}$$

The key packages used in this work include AutoGluon (version 0.4.2),[430] and scikit-learn (version 1.0). For the AutoGluon package, the TabularPredictor function was run with the optional arguments: num_bag_folds=5, num_bag_sets=1, num_stack_levels=1 which has been reported to increase model performance.[430] All code for this study can be found at https://github.com/Goodman-lab/DART-project. The datasets and code are also available on Apollo at https://doi.org/10.17863/CAM.97035 and https://doi.org/10.17863/CAM.100025.

4.3. Results and discussion

The new DART database offers several advantages as compared to using other databases. Firstly, the database contains data that has been published in recent years which ensures that the predictions made with this database are in line with the current DART situation. Secondly, the database is versatile as it offers multiple sources of both *in vivo* and *in vitro* data, as well as developmental toxicity and reproductive toxicity in a single database. This means that one can either use the database for the general prediction of DART or use specific data such as for developmental toxicity only. Finally, this new database incorporates a large number of chemicals that are related to DART, which is more that the popular DART dataset by Wu et al.[377] Modelling with this DART database is thus likely to produce good results. While it is unlikely that the DART database includes all possible chemicals that are related to DART in the literature, the current DART database probably covers enough of the DART compound space to be useful for developing new DART models. It is also noted that as salts and metals are omitted from the DART database, it is expected that predicting the toxicity of compounds containing these inorganics will not work.

4.3.1 Benchmark testing

Before a new package can be applied to the new DART database, a benchmark test must be carried out to verify if the package works as intended. Thus, a benchmark test of the AutoGluon package was carried out with the same data that Feng et al used.[368] In their study, Feng et al developed an ensemble machine learning (ML) model for the specified dataset. They have also compared the performance of their model to a previous study by Jiang et al that used the same dataset.[107, 368] Table 11 summarises the results of their models as well as the models produced by the AutoGluon package on the test set.

Table 11: Test performance of ML models on the dataset used by Feng et al (benchmark)

Model name	ACC (%)	SE (%)	SP (%)	MCC	Ref.
MACCSFP-SVM	83.6	78.5	88.1	-	Jiang et al 2019[107]
Ensemble-Top12	84.4	77.3	90.7	-	Feng et al 2021[368]

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

RandomForestGini_BAG_L2	84.9 ± 3.3	79.6 ± 5.0	89.6 ± 3.8	0.698 ± 0.068	Present study
WeightedEnsemble_L2	85.2 ± 3.4	80.7 ± 5.5	89.2 ± 3.3	0.704 ± 0.069	Present study
WeightedEnsemble_L3	85.6 ± 3.0	79.1 ± 4.4	91.4 ± 3.6	0.714 ± 0.063	Present study
XGBoost_BAG_L1	83.4 ± 3.5	79.3 ± 5.3	87.0 ± 2.8	0.666 ± 0.071	Present study
XGBoost_BAG_L2	84.9 ± 2.9	79.2 ± 3.9	90.0 ± 4.1	0.698 ± 0.062	Present study

Results from the present study are an average of five runs with the data shuffled before each run. The best performing model in the present study is highlighted.

The results in Table 11 demonstrate that the benchmark performance of the models trained by the AutoGluon package in this study is comparable to the models in the literature. In addition, the consistent results of the models across all runs indicate that the AutoML models seem to have been optimised fully as evidenced by the low standard deviations. The high MCC value for all models also suggests that the models are learning well on the data and the results produced are unlikely to be a result of overfitting. The possibility of overfitting is also minimised because the AutoGluon package also has measures to reduce overfitting in their implementation. This could also be due to the consistent method used to train the models using AutoGluon. The documentation of AutoGluon has more details about this and will not be elaborated on in this study.[430] The best performing model produced in this study on the ML dataset was determined to be the WeightedEnsemble. This model has produced an accuracy of 85.6 ± 3.0% which is 2.0% better and 1.2% better than the models developed by Jiang et al and Feng et al respectively. If the standard deviations (taken from an average of five runs using the same parameters) are considered, the model produced by this study is comparable to that produced by Feng et al.[368] One advantage of using AutoGluon is that the model training is extremely quick and without the need for manual hyperparameter optimisation, with the model results for all runs per dataset in the present study being produced within a day. A specified maximum time limit per run can also be specified in the package if necessary. However, it is unlikely that the model training time will exceed this limit if the dataset size is small. The quick model development time, together with the excellent performance metrics comparable to the literature highlights the potential of the AutoGluon package and the use of AutoML in toxicology.

4.3.2 Testing AutoML on datasets

Given the success of the benchmark of the AutoGluon models against known toxicity models, the AutoGluon package was then applied for the DART database where the threshold used for the overall toxicity of each compound is when the overall toxicity value is positive (1) if any of the recorded sources are positive (1). Similarly, a total of 24 models were trained on the DART database across five runs. The metrics for model performance on this dataset are reported in Table 12.

Table 12: Test performance of machine learning models on the DART database

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

Results from the present study are an average of five runs with the data shuffled before each run. The best performing model in the present study is highlighted.

The ExtraTreesClassifier with the 'entropy' criteria produced the best results on the DART database. Similar to the previous benchmark dataset, all models produce similar results and perform consistently well. One would have expected that because the DART endpoint is complex in nature and covers multiple mechanisms/mode of actions, the performance metrics of different ML models would vary more because they generalise the data differently. It is also known that the benchmark dataset only covers reproductive toxicity while the DART database in Table 12 covers both the developmental and reproductive endpoints. This has thus resulted in a lower observed accuracy for all models. Even so, an accuracy of 72.6 ± 0.7% is respectable for modelling the overall toxicity for such a complex endpoint. However, it is acknowledged that when talking about model accuracy, it is ideal to consider the experimental error of the assay. This is because the experimental error would determine the baseline accuracy of the model. It is also observed from Table 11 and Table 12 that varying the algorithm or machine learning model is unlikely to have significant improvements in the performance of the resulting model. It is thus determined that improvements to model performance must come from the perspective of the data used for the models. However, one should also note that it is entirely possible that the data cannot be improved on ie. the experimental error cannot be reduced.

Data from the Stemina assay on a variety of ToxCast chemicals provided in the work by Zurlinden et al has were used for modelling with the AutoGluon package in this thesis chapter.[427] The Stemina assay is an *in vitro* assay for developmental toxicity and measures relative changes in 2 metabolites, ornithine and cystine, targeting the ornithine/cystine ratio as a biomarker for developmental toxicity.[427] Hence, it is useful to investigate how the use of the data from this assay would affect model performance. Table 13 summarises the performance metrics of all ML models trained on this Stemina dataset.

Table 13: Test performance of ML models on the Stemina dataset

Model name	ACC (%)	SE (%)	SP (%)	MCC
CatBoost_BAG_L1	65.6 ± 3.3	25.8 ± 8.0	91.0 ± 5.7	0.232 ± 0.071
CatBoost_BAG_L2	66.5 ± 4.1	29.3 ± 6.1	90.4 ± 4.4	0.257 ± 0.090
ExtraTreesEntr_BAG_L1	66.6 ± 3.3	36.3 ± 6.4	86.0 ± 3.7	0.260 ± 0.074

ExtraTreesEntr_BAG_L2	66.2 ± 4.1	30.7 ± 7.9	89.1 ± 2.5	0.246 ± 0.094
ExtraTreesGini_BAG_L1	66.2 ± 3.2	36.8 ± 6.6	85.0 ± 2.5	0.251 ± 0.079
ExtraTreesGini_BAG_L2	67.7 ± 4.6	33.4 ± 8.9	89.8 ± 4.3	0.287 ± 0.105
LightGBMLarge_BAG_L1	65.8 ± 2.2	28.6 ± 10.4	89.8 ± 4.5	0.238 ± 0.050
LightGBMLarge_BAG_L2	65.2 ± 2.1	26.3 ± 8.5	90.1 ± 3.5	0.216 ± 0.056
LightGBMXT_BAG_L1	64.0 ± 4.4	29.2 ± 7.5	86.4 ± 7.0	0.198 ± 0.074
LightGBMXT_BAG_L2	64.8 ± 2.5	28.7 ± 6.2	88.0 ± 4.0	0.210 ± 0.072
LightGBM_BAG_L1	64.0 ± 4.4	29.2 ± 7.5	86.4 ± 7.0	0.198 ± 0.074
LightGBM_BAG_L2	64.6 ± 2.2	24.0 ± 7.7	90.6 ± 4.8	0.200 ± 0.051
NeuralNetFastAI_BAG_L1	60.9 ± 2.4	52.6 ± 7.9	65.9 ± 5.3	0.184 ± 0.054
NeuralNetFastAI_BAG_L2	59.4 ± 4.5	40.0 ± 9.3	72.0 ± 5.0	0.124 ± 0.102
NeuralNetTorch_BAG_L1	63.1 ± 2.3	30.7 ± 6.2	83.9 ± 5.6	0.176 ± 0.048
NeuralNetTorch_BAG_L2	63.5 ± 4.5	34.4 ± 4.7	82.2 ± 6.2	0.193 ± 0.086
RandomForestEntr_BAG_L1	66.4 ± 3.3	37.1 ± 6.8	85.2 ± 3.3	0.256 ± 0.074
RandomForestEntr_BAG_L2	65.8 ± 4.6	29.6 ± 8.5	89.1 ± 4.3	0.237 ± 0.105
RandomForestGini_BAG_L1	65.7 ± 2.9	36.1 ± 5.9	84.7 ± 2.3	0.238 ± 0.071
RandomForestGini_BAG_L2	66.3 ± 4.2	31.6 ± 8.7	88.6 ± 4.4	0.249 ± 0.101
WeightedEnsemble_L2	65.8 ± 3.3	32.4 ± 8.9	87.3 ± 5.7	0.241 ± 0.070
WeightedEnsemble_L3	63.1 ± 4.7	29.5 ± 5.8	84.6 ± 6.9	0.174 ± 0.086
XGBoost_BAG_L1	65.5 ± 4.8	34.0 ± 8.2	85.8 ± 4.8	0.235 ± 0.098
XGBoost_BAG_L2	63.6 ± 1.9	30.1 ± 6.4	85.2 ± 5.2	0.186 ± 0.028

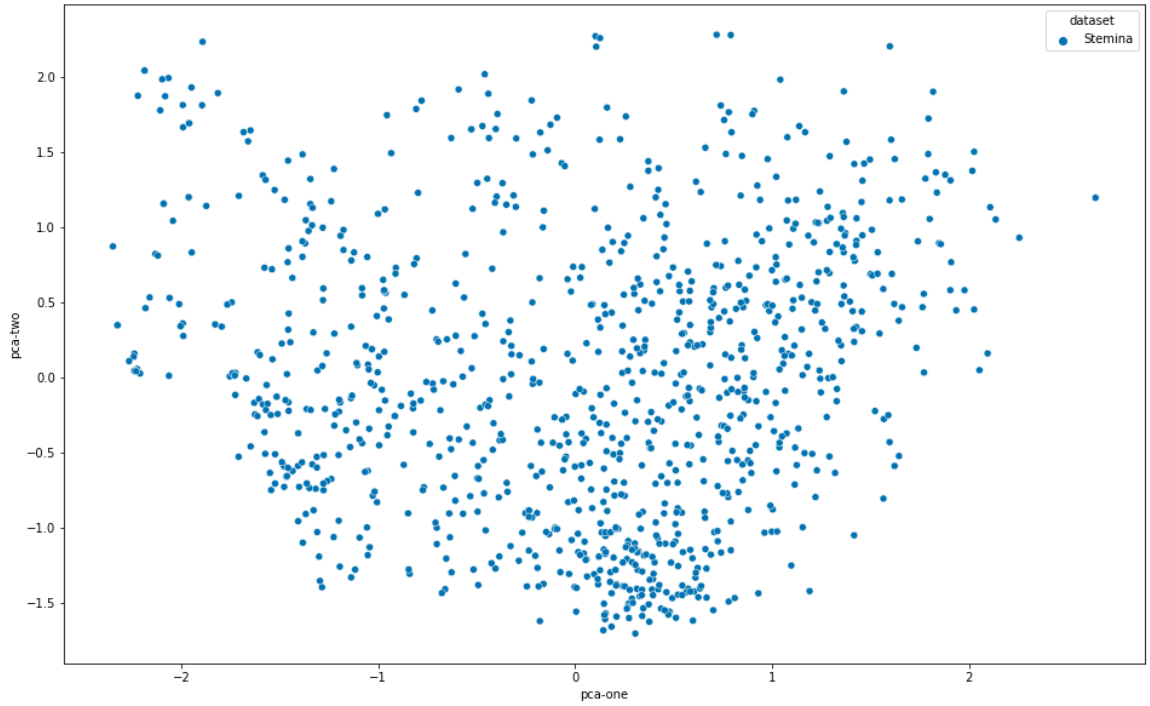
Results from the present study are an average of five runs with the data shuffled before each run. The best performing model in the present study is highlighted.

In Table 13, it is observed that the best performing model is the ExtraTreesClassifier ('Gini impurity' criteria) with an accuracy of 67.7 ± 4.6%. The accuracy is about 5% lower than the best performing model on the DART database. Possible reasons for this observation are that the Stemina data contains highly diverse chemicals or has a low number of similar chemicals that are toxic/positive. As reported by Zurlinden et al, the Stemina dataset contains 19% out of a total of 1065 chemicals that tested positive for developmental toxicity.[427] It is thus likely that the previous reasons hold true as the compounds that test negative are likely to be more dissimilar. The low SE and high SP values are also a result of the class imbalance in the dataset. It is unlikely that the low performance metrics observed in Table 13 are a result of the models or the AutoML process given the two successful cases in Table 11 and Table 12. Class weightage was not used here since the goal is not to produce the best model on this dataset but

to use the same AutoML process and parameters so that all results produced in this study can be compared. In addition, the Stemina data from this dataset is included in the DART database.

Therefore, a principal component analysis (PCA) algorithm as well as a t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm were applied to the Stemina dataset to visualise the feature space. t-SNE is known to be one of the state-of-the-art techniques for dimensionality reduction and feature space visualisation.[431] The resulting plots are shown in Figure 20 and Figure 21. By observing Figure 20 and Figure 21, the feature space of the Stemina dataset does not contain large clusters of data which would be beneficial for modelling. This corroborates with the low performance of machine learning models on the Stemina dataset since machine learning models aim to generalise the given data. This is also evidence that the high proportion of negatives in the dataset are dissimilar to one another.

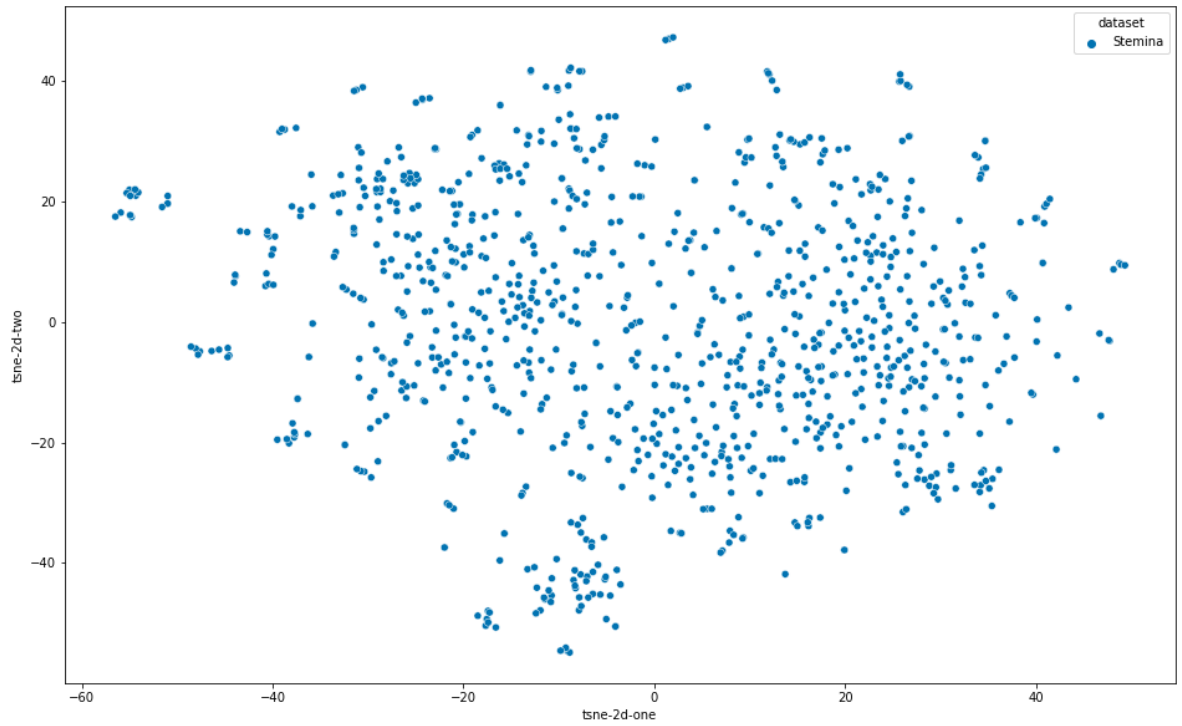
Figure 20: PCA plot for the Stemina dataset



Given the lower performance of the models on the Stemina data, it was hence hypothesised that excluding these data would improve model performance on the DART database. The obtained results are shown in Table 14. From Table 14, it is observed that the best performing model has an accuracy of 78.9 ± 0.5%. As compared to the results in Table 13, the model results obtained demonstrate that it is much better to exclude the 546 unique compounds from the

Stemina dataset, though this will lower the coverage of the model since there are less data points used.

Figure 21: t-SNE plot (perplexity=-20) for the Stemina dataset



Also, the Stemina data uses the ToxCast library which covers environmental and non-drug like chemicals while is different to most of the other data sources. From a modelling perspective, it would be better to exclude the Stemina data when modelling general DART as a toxicological endpoint, however the Stemina data can be included when modelling more specific endpoints that is related to teratogenicity as the Stemina assay measures teratogenicity potential.

Table 14: Test performance of machine learning models on the DART database excluding the compounds in the Stemina dataset

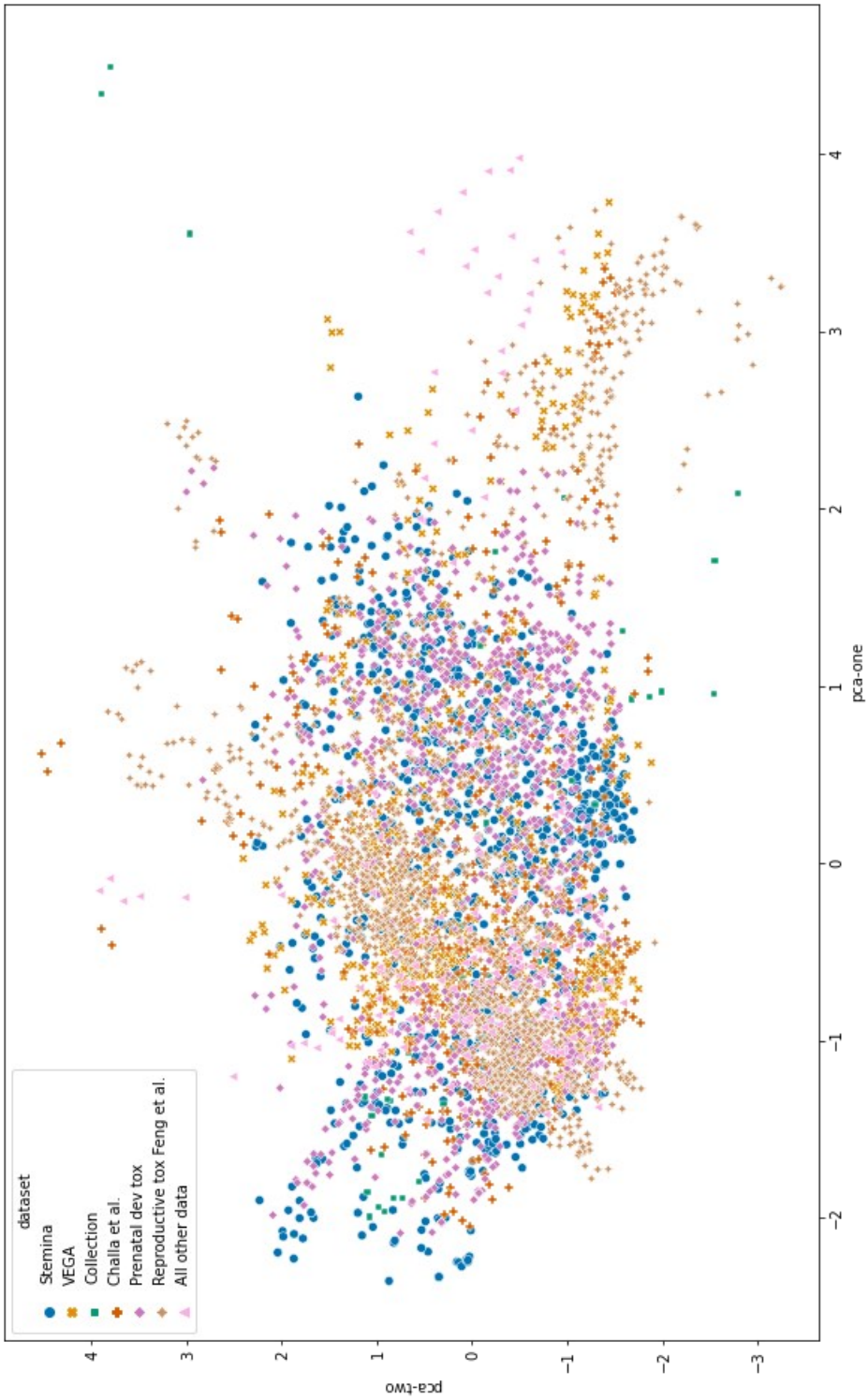
Model name	ACC (%)	SE (%)	SP (%)	MCC
CatBoost_BAG_L1	77.1 ± 0.9	81.3 ± 2.9	71.6 ± 4.0	0.531 ± 0.017
CatBoost_BAG_L2	78.5 ± 1.4	82.9 ± 2.7	72.8 ± 4.3	0.560 ± 0.025
ExtraTreesEntr_BAG_L1	78.2 ± 1.3	82.2 ± 2.9	72.8 ± 2.8	0.553 ± 0.023
ExtraTreesEntr_BAG_L2	78.7 ± 0.8	82.6 ± 3.0	73.5 ± 3.4	0.564 ± 0.013
ExtraTreesGini_BAG_L1	78.6 ± 0.9	84.1 ± 3.0	71.3 ± 4.0	0.561 ± 0.015
ExtraTreesGini_BAG_L2	78.8 ± 0.9	83.1 ± 2.3	73.2 ± 4.4	0.566 ± 0.017

LightGBMLarge_BAG_L1	77.0 ± 1.3	81.7 ± 3.5	71.0 ± 7.4	0.531 ± 0.030
LightGBMLarge_BAG_L2	77.6 ± 1.2	82.1 ± 3.1	71.8 ± 4.4	0.542 ± 0.022
LightGBMXT_BAG_L1	77.1 ± 1.9	81.7 ± 2.9	71.2 ± 6.6	0.532 ± 0.042
LightGBMXT_BAG_L2	78.1 ± 1.1	83.7 ± 3.2	70.8 ± 4.9	0.552 ± 0.020
LightGBM_BAG_L1	77.1 ± 1.9	81.7 ± 2.9	71.2 ± 6.6	0.532 ± 0.042
LightGBM_BAG_L2	77.6 ± 2.1	83.0 ± 3.2	70.6 ± 4.4	0.542 ± 0.038
NeuralNetFastAI_BAG_L1	76.4 ± 2.3	80.5 ± 4.1	70.7 ± 5.0	0.515 ± 0.046
NeuralNetFastAI_BAG_L2	78.4 ± 1.9	82.1 ± 1.8	73.3 ± 4.3	0.556 ± 0.041
NeuralNetTorch_BAG_L1	77.6 ± 1.2	82.2 ± 1.9	71.3 ± 3.3	0.539 ± 0.030
NeuralNetTorch_BAG_L2	78.9 ± 0.5	80.9 ± 2.6	76.4 ± 2.5	0.571 ± 0.009
RandomForestEntr_BAG_L1	78.7 ± 1.1	82.7 ± 3.3	73.5 ± 3.2	0.564 ± 0.018
RandomForestEntr_BAG_L2	78.4 ± 0.9	82.8 ± 2.6	72.5 ± 3.8	0.557 ± 0.015
RandomForestGini_BAG_L1	78.1 ± 1.3	83.4 ± 3.2	71.2 ± 3.8	0.552 ± 0.021
RandomForestGini_BAG_L2	78.8 ± 1.2	82.8 ± 2.8	73.5 ± 3.6	0.566 ± 0.023
WeightedEnsemble_L2	78.4 ± 1.0	82.8 ± 3.0	72.5 ± 3.6	0.557 ± 0.018
WeightedEnsemble_L3	78.3 ± 1.3	82.6 ± 3.1	72.7 ± 4.4	0.556 ± 0.023
XGBoost_BAG_L1	77.3 ± 1.1	80.5 ± 4.0	73.2 ± 5.2	0.538 ± 0.025
XGBoost_BAG_L2	77.6 ± 1.7	81.6 ± 3.5	72.3 ± 4.2	0.542 ± 0.031

Results from the present study are an average of five runs with the data shuffled before each run. The best performing model in the present study is highlighted.

The feature space of all the chemicals in the DART database has also been investigated. PCA and t-SNE plots of the DART database separated by the data source have been constructed and shown in Figure 22 and Figure 23. Based on Figure 22 and Figure 23, the feature space of the different data sources seems to be within the same region. Some local clustering is also observed for the t-SNE plot which suggests similarity between some groups of chemicals, especially that the data points for "All other data" are structurally distinguished from rest of the data points. This is probably because this set of data points contains many compounds targeting very specific targets/endpoints including thyroid bioactivity, androgen receptor, and steroidogenesis.

Figure 22: PCA plots for the DART database separated by the data source



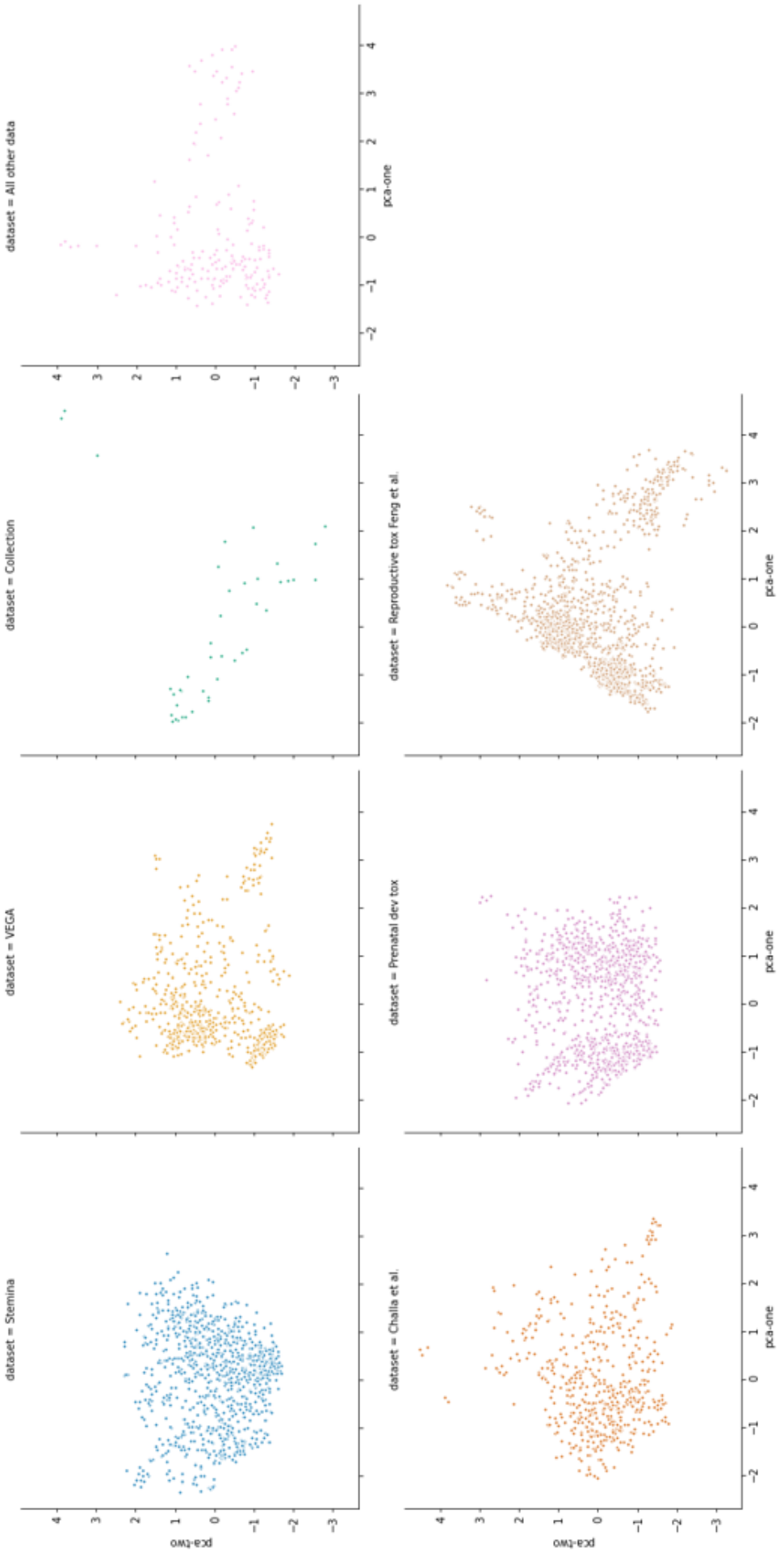
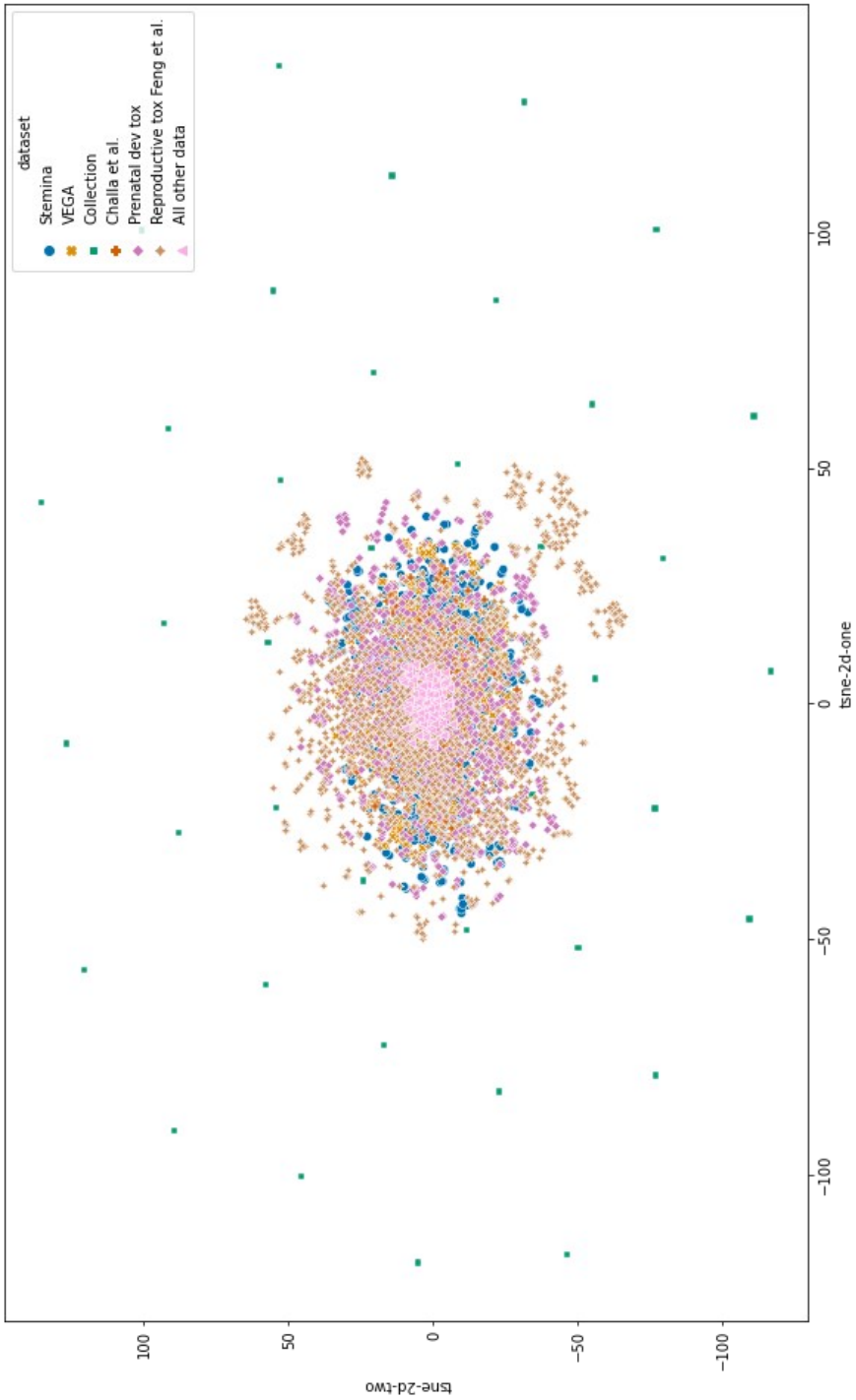

dataset = All other data
dataset = Collection
dataset = VEGA
dataset = Reproductive tox Feng et al.
dataset = Prenatal dev tox
dataset = Stemina
dataset = Challa et al.
pca-one
pca-two

Figure 23: t-SNE plot (perplexity=40) for the DART database separated by the data source



4.3.3 Testing on an external test set

Moving on, the models trained on the DART database applied to the external test set. This dataset was obtained from Hewitt et al and contains 290 compounds for developmental toxicity.[429] As compared to reproductive toxicity which has already been benchmarked previously, a similar benchmark has not been done for the models in the present study on

developmental toxicity. It is thus worthwhile to investigate the performance of the models in the present study on compounds that are related to developmental toxicity but not present in the DART database. This external test dataset was processed in a similar fashion to all the other datasets used in this study. After processing, the external test dataset contains a total of 68 chemicals (41 positives, 27 negatives). The performance metrics of the ML models developed in this study on this external test set are tabulated in Table 15.

Table 15: Test performance of ML models in Table 12 on an external test set

Model name	ACC (%)	SE (%)	SP (%)	MCC
CatBoost_BAG_L1	73.5 ± 2.1	79.5 ± 2.2	64.4 ± 2.0	0.443 ± 0.043
CatBoost_BAG_L2	77.6 ± 2.6	82.4 ± 6.5	70.4 ± 3.7	0.534 ± 0.045
ExtraTreesEntr_BAG_L1	78.5 ± 0.8	87.3 ± 3.2	65.2 ± 4.2	0.546 ± 0.016
ExtraTreesEntr_BAG_L2	76.5 ± 3.1	82.9 ± 5.5	66.7 ± 2.6	0.505 ± 0.062
ExtraTreesGini_BAG_L1	80.0 ± 2.5	88.8 ± 4.4	66.7 ± 2.6	0.578 ± 0.054
ExtraTreesGini_BAG_L2	77.1 ± 1.3	84.9 ± 3.6	65.2 ± 3.3	0.515 ± 0.026
LightGBMLarge_BAG_L1	75.9 ± 6.1	82.4 ± 9.4	65.9 ± 1.7	0.496 ± 0.122
LightGBMLarge_BAG_L2	72.4 ± 3.0	80.0 ± 4.7	60.7 ± 2.0	0.416 ± 0.059
LightGBMXT_BAG_L1	75.9 ± 2.9	80.0 ± 4.7	69.6 ± 3.1	0.498 ± 0.055
LightGBMXT_BAG_L2	76.8 ± 1.9	82.9 ± 6.7	67.4 ± 6.6	0.515 ± 0.038
LightGBM_BAG_L1	75.9 ± 2.9	80.0 ± 4.7	69.6 ± 3.1	0.498 ± 0.055
LightGBM_BAG_L2	73.5 ± 2.8	79.5 ± 4.8	64.4 ± 3.3	0.445 ± 0.053
NeuralNetFastAI_BAG_L1	71.5 ± 3.8	72.2 ± 6.8	70.4 ± 2.6	0.421 ± 0.066
NeuralNetFastAI_BAG_L2	72.9 ± 4.1	74.6 ± 8.2	70.4 ± 5.9	0.448 ± 0.074
NeuralNetTorch_BAG_L1	74.4 ± 4.0	79.0 ± 6.6	67.4 ± 3.1	0.467 ± 0.073
NeuralNetTorch_BAG_L2	76.5 ± 2.3	82.0 ± 5.6	68.1 ± 5.6	0.508 ± 0.043
RandomForestEntr_BAG_L1	80.3 ± 2.5	89.3 ± 4.8	66.7 ± 5.2	0.585 ± 0.053
RandomForestEntr_BAG_L2	74.1 ± 3.8	80.0 ± 7.4	65.2 ± 3.3	0.459 ± 0.069
RandomForestGini_BAG_L1	79.1 ± 1.2	87.8 ± 3.0	65.9 ± 1.7	0.558 ± 0.027
RandomForestGini_BAG_L2	74.7 ± 3.2	81.5 ± 6.1	64.4 ± 3.3	0.468 ± 0.058
WeightedEnsemble_L2	78.2 ± 2.4	85.4 ± 3.9	67.4 ± 1.7	0.54 ± 0.05
WeightedEnsemble_L3	73.8 ± 2.8	79.5 ± 6.1	65.2 ± 4.2	0.453 ± 0.052
XGBoost_BAG_L1	74.4 ± 2.7	77.1 ± 2.8	70.4 ± 5.2	0.471 ± 0.057
XGBoost_BAG_L2	73.8 ± 5.0	79.5 ± 8.0	65.2 ± 2.0	0.454 ± 0.099

Results from the present study are an average of five runs with the data shuffled before each run. The best performing model in the present study is highlighted.

Surprisingly, the performance of the ML models in the present study is higher on the external test set. This could mean that the compounds in this external test set are largely similar to those in the DART database; that is the external test set is within the models' applicability domain. Also, the consistent model performance of all models shows that the models developed in this study are viable for the general prediction of DART.

4.3.4 Developmental toxicity vs. reproductive toxicity

As compared to reproductive toxicity (1606 chemicals, 848 positives, 758 negatives), data for developmental toxicity covers a larger chemical space (2202 chemicals, 1206 positives, 996 negatives). About 600 compounds are common between both endpoints, which corresponds to about 27% and 38% for developmental toxicity and reproductive toxicity respectively. Thus, the PCA and t-SNE plots of the DART database separated by the toxicity endpoint have been constructed and shown in Figure 24 and Figure 25. Data solely from Wu et al was omitted as the data source classifies chemicals for both endpoints and it would not affect the overall plot visualisation significantly.[377] Also, the goal is not to quantify the feature space for both endpoints but to get a qualitative visual representation of the feature space of DART.

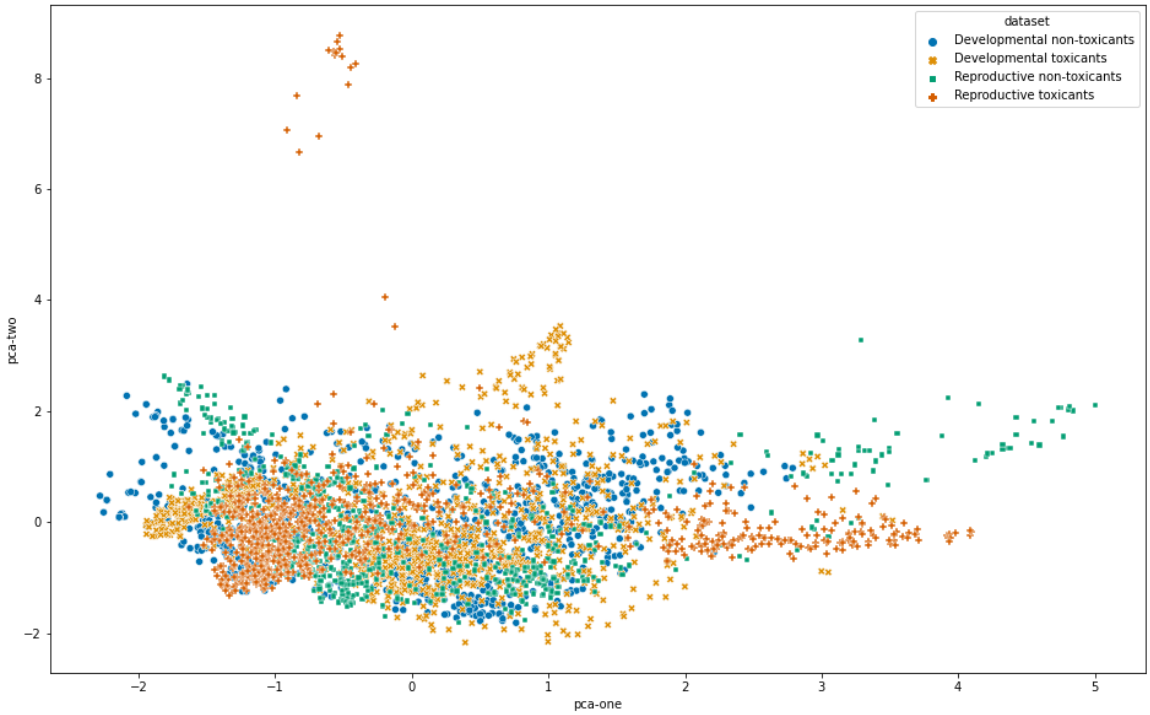
Figure 24 and Figure 25 indeed shows that the toxicants for reproductive toxicity are clustered closer together as compared to those for developmental toxicity. This is probably because the reproductive toxicity data in the DART database is made up of more specific endpoints as compared to developmental toxicity. The plots also suggest that more data on developmental toxicity is likely to improve model performance since the data for developmental toxicity covers a larger feature space. ML models were also trained separately on the endpoints with the results shown in Table 16 where it is observed that the models for developmental toxicity have lower performance metrics than reproductive toxicity. This is likely because of the Stemina data where the ML models have been observed to have low performance (Table 13) as well as the data from Ciallella et al on prenatal developmental toxicity which also had low model performance.[365] This suggests that more data is needed in order to improve the predictive power for developmental toxicity models.

Table 16: Test performance of the best-performing ML models for developmental toxicity and reproductive toxicity separately

Toxicity endpoint	Model name	ACC (%)	SE (%)	SP (%)	MCC
Developmental	WeightedEnsemble_L2	67.6 ± 2.9	75.7 ± 4.1	57.6 ± 3.7	0.340 ± 0.062
Reproductive	RandomForestEntr_BAG_L1	80.4 ± 1.5	79.7 ± 2.4	81.3 ± 2.8	0.609 ± 0.031

Results from the present study are an average of five runs with the data shuffled before each run.

Figure 24: PCA plot for the DART database separated by the toxicity endpoint



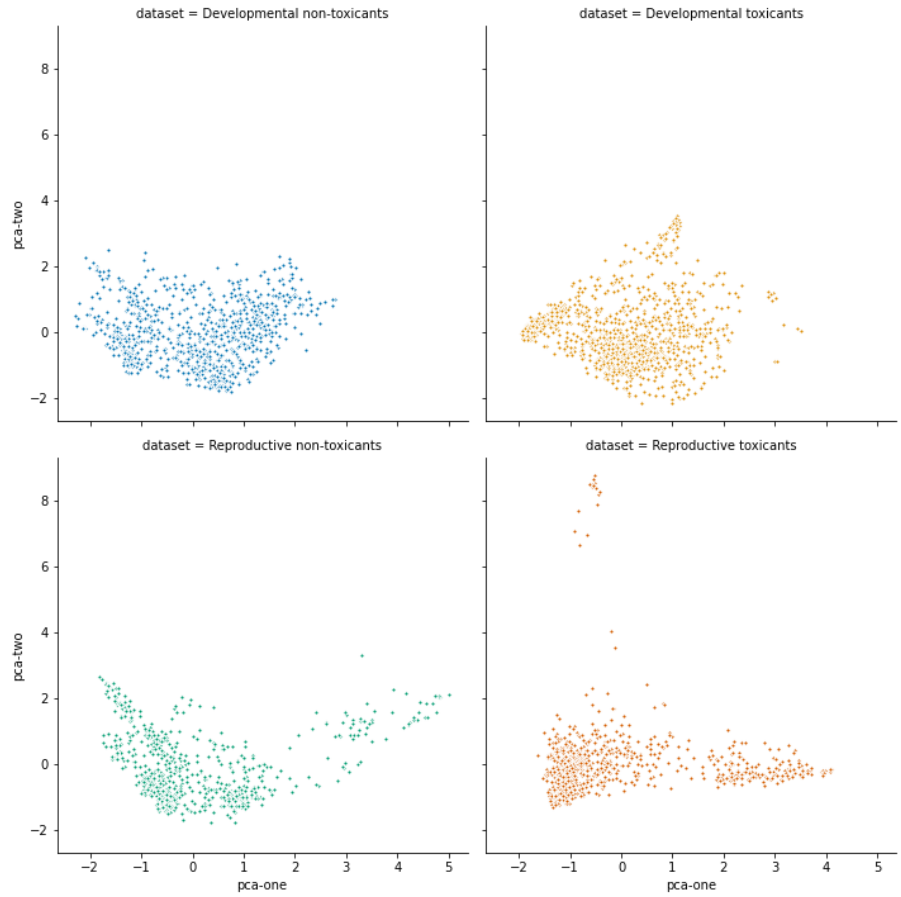


Figure 25: t-SNE plot (perplexity=10) for the DART database separated by the toxicity endpoint

4.3.5 *in vivo* vs. *in vitro*

In section 4.3.2, the performances of ML models on the Stemina dataset (*in vitro* data) were reported (Table 13). In this section, a more general investigation was done on all *in vivo* data (1621 chemicals, 878 positives, 743 negatives) vs. all *in vitro* data (1064 chemicals, 439 positives, 625 negatives) according to the classification listed in the DART database. As there is a larger number of *in vivo* data, it is expected that a larger chemical space and thus a broader spectrum of the mechanism of action is covered. Once again, feature plots were constructed with the results shown in Figure 26 and Figure 27.

Figure 26 and Figure 27 show that the two test types cover a different feature space and could just be due to the fact that the *in vivo* data contains more chemicals than *in vitro* data. The regulatory assessment of carcinogenicity and DART still relies heavily on animal testing but this is not the case for cosmetics where animal tests are banned.[369] It is clear that more work, in particular adverse outcome pathways, needs to be done in this field to develop non-animal alternatives for the DART endpoint as well as to gain acceptance of the current *in vitro* methods.

Figure 26: PCA plot for the DART database separated into *in vivo* vs. *in vitro* data

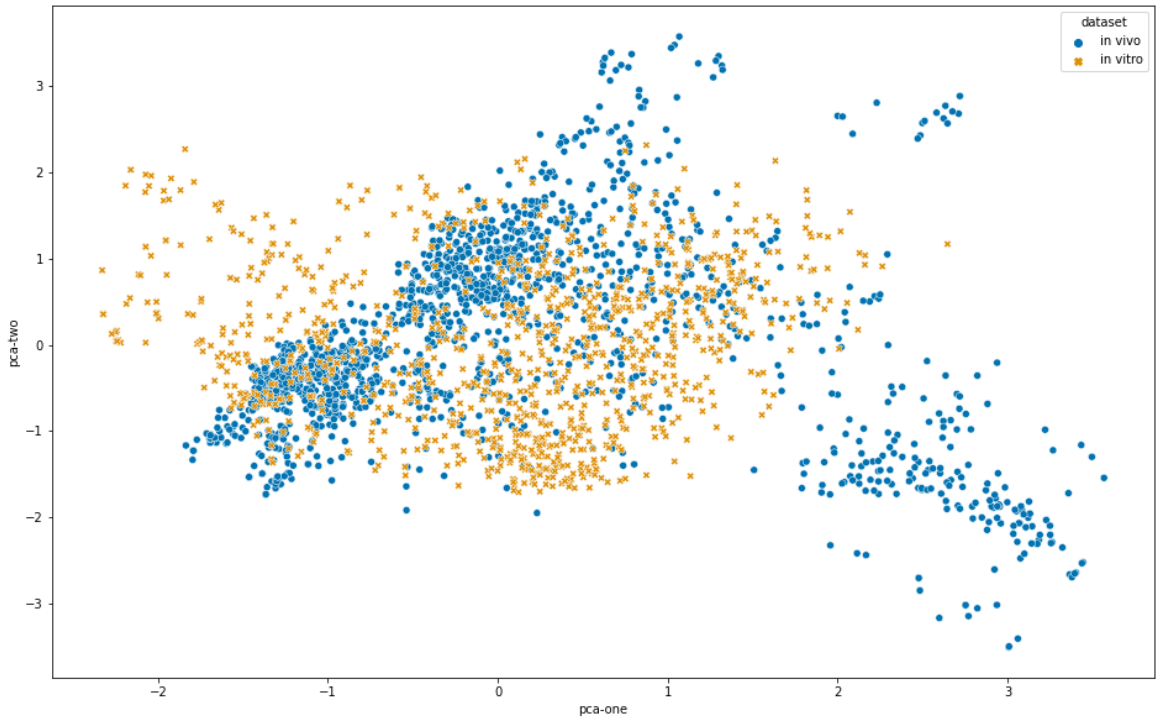
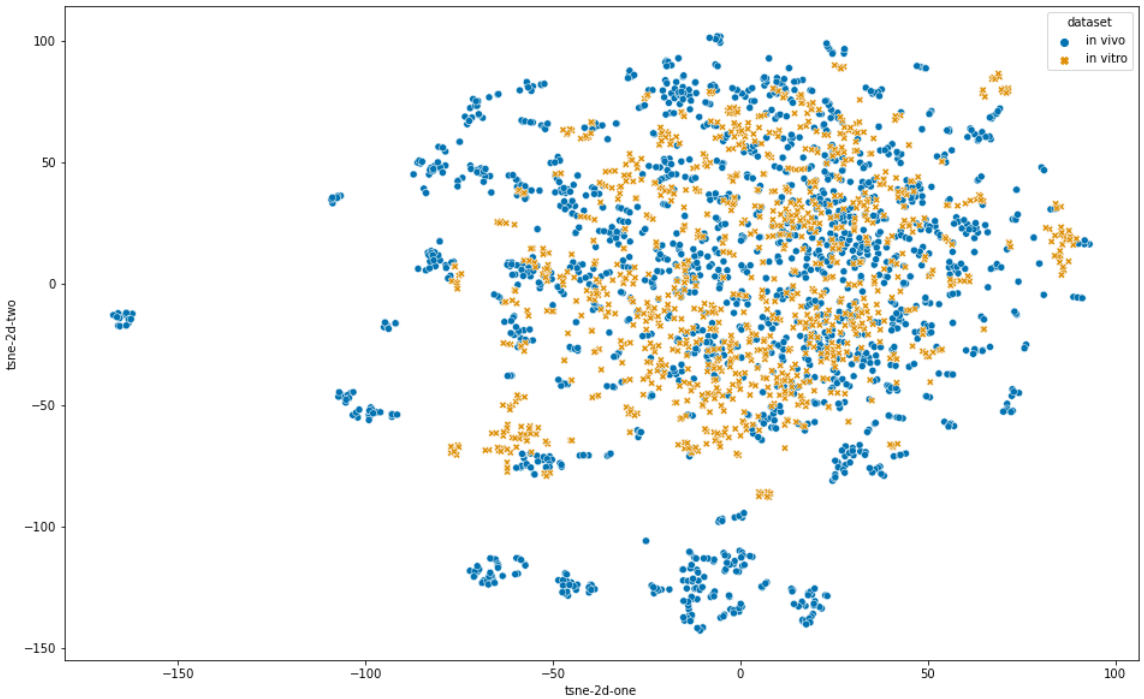


Figure 27: t-SNE plot (perplexity=10) for DART database separated into *in vivo* vs. *in vitro* data



ML models have also been trained separately on *in vivo* and *in vitro* data. These results are tabulated in Table 17. As evidenced in Table 17, the best-performing model trained on *in vivo* data has higher performance metrics than that for *in vitro* data. This is likely because the main contributor to *in vivo* data is the data taken from Feng et al which has good model performance (Table 11) while the main contributor to *in vitro* data is the data taken from the work by Zurlinden et al which has been shown to have poor model performance (Table 13).[368, 427]

Table 17: Test performance of the best-performing ML models for the test types (*in vivo* versus *in vitro*)

Test type	Model name	ACC (%)	SE (%)	SP (%)	MCC
in vivo	RandomForestEntr_BAG_L2	82.0 ± 0.7	79.1 ± 2.7	85.6 ± 4.7	0.644 ± 0.023
in vitro	RandomForestGini_BAG_L2	64.2 ± 3.8	30.5 ± 6.0	89.2 ± 2.0	0.246 ± 0.070

Results from the present study are an average of five runs with the data shuffled before each run.

4.3.6 Varying thresholds for the overall toxicity value

A different threshold for the overall toxicity value was also used to investigate if varying the positive threshold condition can increase SP and lower SE, which could be applied for screening purposes. As the original threshold produced a high SE, a different threshold would be required if high SP was desired. The threshold for positives for the overall toxicity value was thus adjusted to being positive only if the number of sources where the compound has been tested to be positive are greater than the number of sources where the compound has been tested to be negative. This results in the number of positives and negatives being 1424 and 1821 respectively as compared to 1662 and 1583 respectively for the original threshold. The results for both thresholds are shown in Table 18. It is observed that the change in the threshold does increase SP as desired and even causes a slight increase in model accuracy. The choice of which threshold to use would depend on the purpose of the user. For example, if the purpose is to screen compounds, the original threshold of "at least one positive" would be favoured as the performance metrics (SE and SP) are more balanced. For toxicity confirmation, the other threshold would be preferred as the best-performing model's SP is higher.

Table 18: Test performance of the best-performing ML models for varying positive thresholds

Threshold for positives	Model name	ACC (%)	SE (%)	SP (%)	MCC
At least one positive	ExtraTreesEntr_BAG_L2	72.6 ± 0.7	71.8 ± 2.9	73.6 ± 2.9	0.453 ± 0.014
Positives more than negatives	CatBoost_BAG_L2	75.1 ± 2.6	67.0 ± 2.8	81.3 ± 4.3	0.491 ± 0.052

Results from the present study are an average of five runs with the data shuffled before each run.

4.3.7 Maximising SE or SP

An alternative approach to maximising SE or SP was taken by adjusting the AutoML parameters. Custom scorer functions for SE or SP were passed to the TabularPredictor function using the eval_metric parameter during model initialization with the new goal of maximising SE or SP during the training/validation process. These results are shown in Table 19.

The results in Table 19 show that the limits for SE and SP are about 75% to 76% using the original threshold. This is not a significant improvement from the general best-performing model for DART (Table 12) that aims to maximise accuracy. Once again, this demonstrates that the performance of ML models on this DART database has reached a limit with further improvements likely requiring a higher quantity and quality of data.

Table 19: Test performance of the best-performing ML models for maximising SE or SP during the training/validation process

Maximise	Model name	ACC (%)	SE (%)	SP (%)	MCC
SE	ExtraTreesGini_BAG_L1	73.3 ± 1.5	75.3 ± 1.9	71.3 ± 4.0	0.467 ± 0.028
SP	ExtraTreesGini_BAG_L2	74.4 ± 1.4	73.3 ± 1.5	75.6 ± 1.9	0.489 ± 0.028

Results from the present study are an average of five runs with the data shuffled before each run. Lesser models were trained as some model algorithms were not suitable for the adjustment of the validation function. A total of 10 ML models were trained.

4.3.8 Consensus approach

In order to improve on the current results for developmental toxicity and reproductive toxicity, a consensus approach where the predictions made by all the models for the endpoint were first recorded, followed by weighting, and finally evaluating the overall prediction based on a threshold was attempted. Models for developmental toxicity and reproductive toxicity were

developed and analysed separately as it would allow for the predictions made by the models to be understood more effectively as compared to a global DART model (using all the data in the DART database for modelling) where the prediction could be due to either endpoint. Figure 28 shows a flowchart representing this idea while the results are shown in Table 20. The consensus prediction in Table 20 was taken to be the majority prediction across all models. For example, if 13 models predict a compound is toxic while 11 models predict a compound is non-toxic, the compound would be classified as toxic. However, one should note that this would mean a low confidence for this prediction as there are 11 models predicting that the compound is non-toxic.

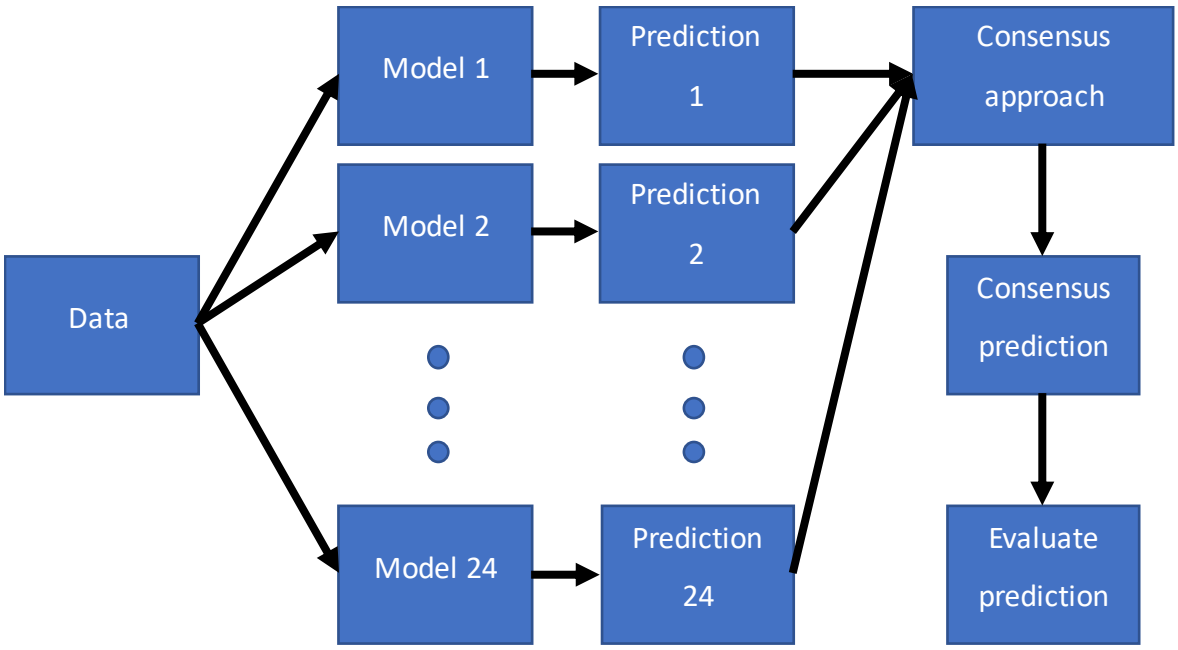


Figure 28: A flowchart showing the consensus approach to determine the overall prediction

From the results in Table 20, it is observed that the consensus approach is comparable to the non-consensus approach which uses a single model, with models for reproductive toxicity having a larger difference in performance metrics between both approaches as compared to that for developmental toxicity. The use of the consensus approach also offers an advantage over the non-consensus approach, where the number of models that agree on the classification of the compound could be used as a level of confidence in the prediction.

In order to further investigate the consensus approach, a strict threshold was applied for the consensus approach where all 24 models must agree on the classification of a compound before

the classification is accepted. Otherwise, the compound's prediction would be treated as uncertain, and the compound will be excluded from the final evaluation. These results are summarised in Table 21.

Table 20: Test performance of the best-performing ML models for developmental toxicity and reproductive toxicity separately using the consensus approach

Toxicity endpoint	Model name	ACC (%)	SE (%)	SP (%)	MCC
Developmental[a]	WeightedEnsemble_L2	67.6 ± 2.9	75.7 ± 4.1	57.6 ± 3.7	0.340 ± 0.062
Developmental[b]	Consensus	67.2	71.0	62.5	0.336
Reproductive[a]	RandomForestEntr_BAG_L1	80.4 ± 1.5	79.7 ± 2.4	81.3 ± 2.8	0.609 ± 0.031
Reproductive[b]	Consensus	77.0	78.2	75.7	0.539

[a] Results from the present study are an average of five runs with the data shuffled before each run.

[b] Results were obtained using the models from run 1 as the performances are comparable across all runs.

Table 21: Test performance of models using the consensus approach using the strict threshold for developmental toxicity and reproductive toxicity separately

Toxicity type	Were rows/data dropped?	No. of rows/data dropped	Accuracy (%)	SE (%)	SE (%)	SP (%)	MCC
Developmental	No	-	67.2	71.0	71.0	62.5	0.336
	Yes	214 (48.1%)	77.1	81.1	81.1	72.1	0.535
Reproductive	No	-	77.0	78.2	78.2	75.7	0.539
	Yes	98 (30.4%)	85.3	87.0	87.0	82.8	0.697

As expected, the removal of compounds whose predictions are uncertain improves the final evaluation. The consensus approach adopted here can thus determine when one should trust the prediction *i.e.* all 24 models agree or be cautious *i.e.* the prediction is uncertain. The improvement in performance is also rather large, with 10.1% and 9.2% for developmental and reproductive toxicity respectively. The number of models used as the majority prediction was also varied, with selected results tabulated in Table 22 and the full results in Table A16 of the

Appendices. Based on the results obtained, the best performance is obtained when the consensus prediction is agreed on by all 24 models though this would reduce the coverage.

As mentioned earlier, the number of rows dropped is also an indication of the confidence of the ML models when predicting the toxicity endpoint. The models are more uncertain about compounds from developmental toxicity as compared to reproductive toxicity as observed from the larger percentage of rows/data dropped. Once again, this reflects the complexity of the endpoint, or it could possibly be due to a lack of quality data for that particular endpoint. A further improvement in model performance as well as the understanding of the data would be likely if more data can be gathered for both developmental toxicity and reproductive toxicity, though this remains a challenge in the field. The models for more specific mechanisms covered by DART should be better than models for general reproductive or developmental effects and the results in this chapter do demonstrate this to an extent.

Table 22: Test performance using the consensus approach with varying consensus thresholds for developmental toxicity and reproductive toxicity separately

Toxicity type	No. of models used for majority prediction	SE (%)	SP (%)	Accuracy (%)	MCC	TP	FP	FN	TN
Developmental	24	81.1	72.1	77.1	0.535	103	29	24	75
	21	74.5	65.8	70.6	0.404	146	54	50	104
	16	72.9	63.8	68.8	0.368	164	68	61	120
	13	70.7	63.1	67.3	0.338	171	73	71	125
Reproductive	24	87.0	82.8	85.3	0.697	114	16	17	77
	21	82.1	80.5	81.4	0.625	124	24	27	99
	16	78.0	77.9	78.0	0.559	128	32	36	113
	13	77.8	76.2	77.0	0.540	130	36	37	115

4.3.9 Misclassification of ML models with the consensus approach

It is also useful to know when the ML models for developmental toxicity and reproductive toxicity are poor at predicting the toxicity of the test compounds for the results obtained using the consensus approach when all models agree on the prediction. Therefore, k-means clustering was used to investigate if there are any similarities between compounds in a cluster for both the developmental toxicity data and the reproductive toxicity data. The optimal number of clusters for k-means was determined to be five using the elbow method (Appendices Figure A16 and Figure A17) for both the developmental toxicity data and the reproductive toxicity data. The number of false positives (FP) and false negatives (FN) in each cluster were thus tabulated in order to identify when the models will misclassify compounds, and this is shown in Table 23. The numbers in Table 23 indicate how many FPs and FNs (misclassified data points) are present in each cluster of the developmental toxicity data or the reproductive toxicity data.

Table 23: Number of false positives and false negatives for developemental toxicity and reproductive toxicity using the consensus approach

Toxicity endpoint	Label	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
Developmental	FP	4	8	4	5	8	29
Developmental	FN	5	7	-	8	4	24
Reproductive	FP	3	4	7	1	1	16
Reproductive	FN	-	7	5	3	2	17

The values shown represent the number of misclassified data points. For example, there are four FP labelled data points in cluster 0 for developmental toxicity. This means that four data points in the developmental toxicity data belonging to cluster 0 have been misclassified and are false positives.

Based on Table 23, there is no cluster with a significant number of misclassified data points, indicating that the models do not misclassify certain groups (based on the clusters) of compounds on a large scale. In order to investigate further, the structures of the compounds in each cluster were drawn and visualized using RDKit. These structures are shown in Figure 29 to Figure 32.

Although the number of structures in certain clusters are limited, it is observed that some conclusions can be made regarding the misclassification of compounds by the models. For

developmental toxicity, the misclassified compounds are mainly long chain alkyl compounds as well as a number of aromatic or sp^2 nitrogen-containing compounds. For reproductive toxicity, long chain alkyl compounds are also misclassified. A possible reason is that these long chain alkyl compounds are surfactants and thus difficult to predict because of their properties, but they have to be tested since they are one of the main ingredients of cosmetics.[432]

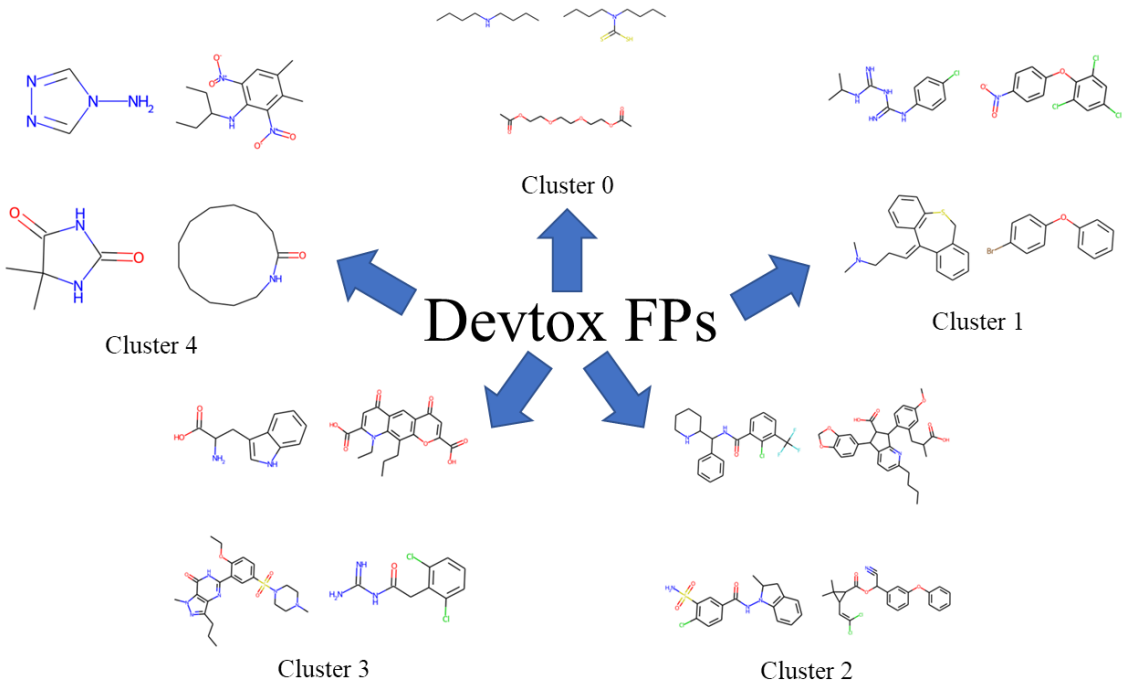


Figure 29: Representative structures for the clusters with the developmental toxicity FP label

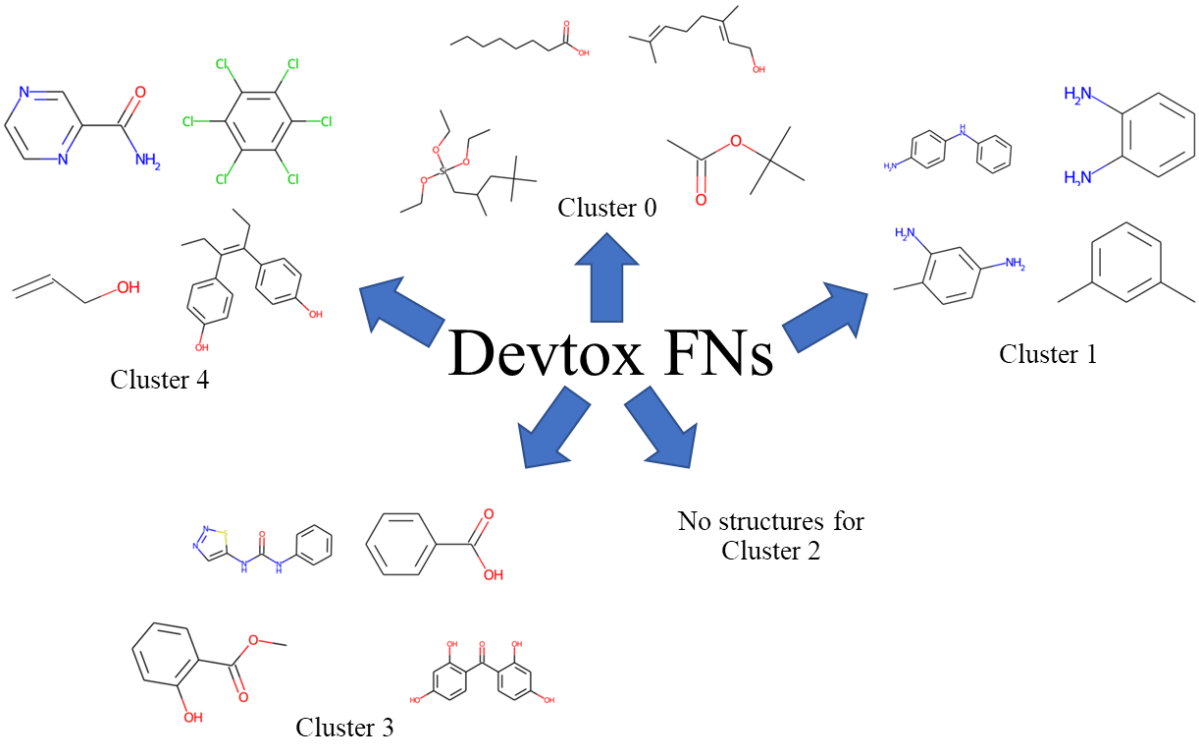


Figure 30: Representative structures for the clusters with the developmental toxicity FN label

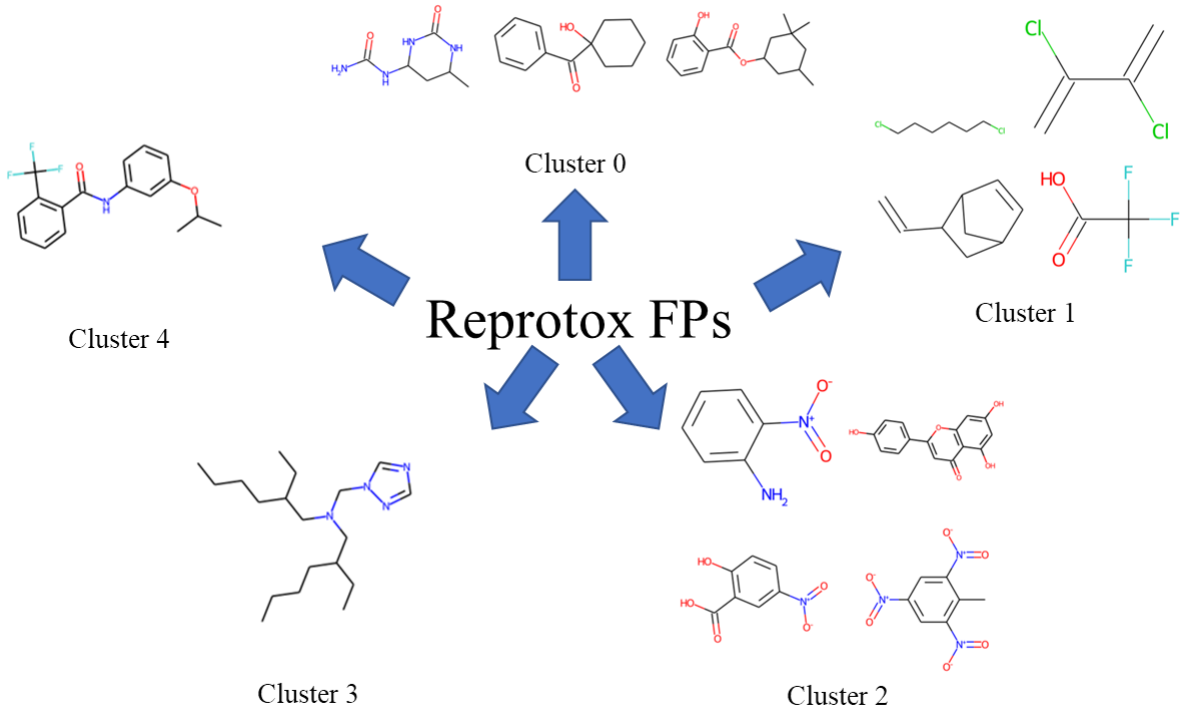


Figure 31: Representative structures for the clusters with the reproductive toxicity FP label

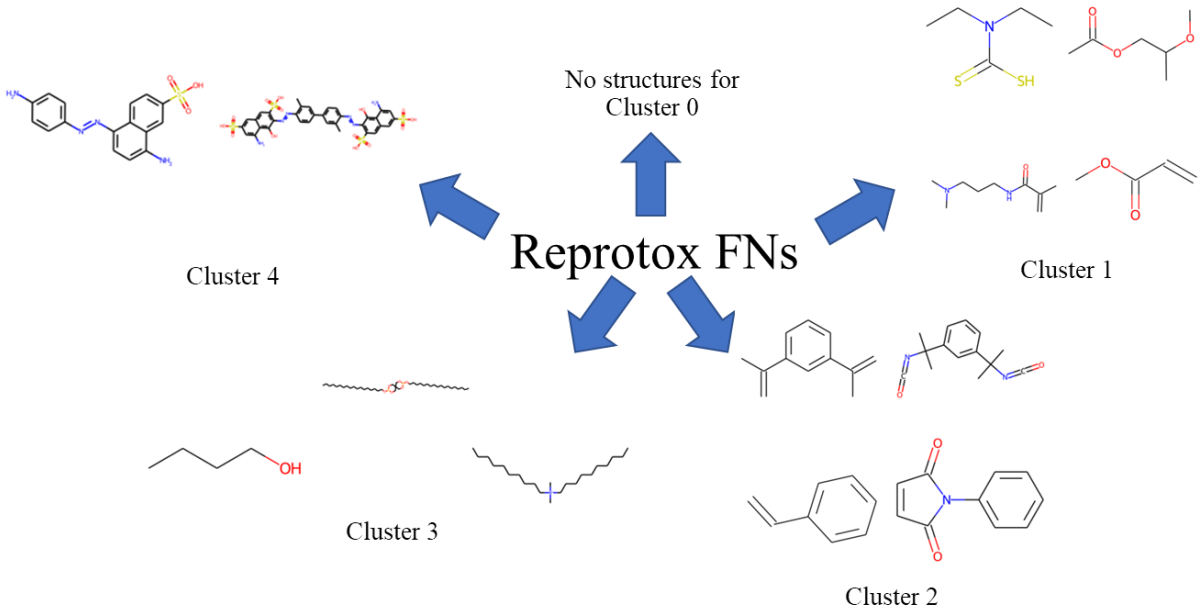


Figure 32: Representative structures for the clusters with the reproductive toxicity FN label

Some clusters for the developmental toxicity data or the reproductive toxicity data also have zero misclassified compounds, indicating that if a new compound falls within that cluster, it would be likely to be classified correctly.

It is noted that the conclusions that can be drawn from the misclassification of data are limited as the number of structures in each cluster are small. To take a cautious approach when convincing regulators on the safety of products, these models can be used together with *in vitro* testing methods in a weight of evidence approach.

4.3.10 Transfer learning between developmental toxicity and reproductive toxicity

In this section, the use of transfer learning between developmental toxicity and reproductive toxicity is investigated. The results are summarised in Table 24 and Table 25 and suggests that the use of transfer learning between the two toxicity endpoints is viable; that is models trained on reproductive toxicity data can be applied to developmental toxicity data and vice versa. The performance of these ML models is also comparable to the ML models trained on the endpoints. According to the European Chemicals Agency (ECHA), reproductive toxicity refers to effects on both fertility and development, but also refers to effects on fertility alone. Developmental toxicity refers to chemicals' interference with normal development of the organism, originating from exposure of either parent prior to conception, or exposure of the developing organism.[433]

Thus, such results suggest that there is an overlap between the two toxicity endpoints, which is in line with the definition of the two toxicity endpoints.

Table 24: Test performance for transfer learning of models developed on reproductive toxicity and used to predict on developmental toxicity data

Model name	ACC (%)	SE (%)	SP (%)	MCC
CatBoost_BAG_L1	63.9 ± 1.0	77.4 ± 1.7	47.2 ± 1.0	0.259 ± 0.021
CatBoost_BAG_L2	67.7 ± 0.5	80.7 ± 0.9	51.5 ± 2.1	0.34 ± 0.010
ExtraTreesEntr_BAG_L1	67.9 ± 0.4	82.4 ± 1.0	50.0 ± 1.8	0.346 ± 0.008
ExtraTreesEntr_BAG_L2	66.6 ± 0.5	80.0 ± 1.3	50.1 ± 2.2	0.317 ± 0.010
ExtraTreesGini_BAG_L1	67.8 ± 0.2	82.5 ± 1.4	49.6 ± 1.7	0.343 ± 0.004
ExtraTreesGini_BAG_L2	66.8 ± 0.3	80.2 ± 1.2	50.2 ± 1.9	0.321 ± 0.007
LightGBMLarge_BAG_L1	67.3 ± 1.0	83.0 ± 1.0	47.9 ± 2.7	0.333 ± 0.021
LightGBMLarge_BAG_L2	67.0 ± 0.5	80.6 ± 1.0	50.3 ± 1.7	0.326 ± 0.011
LightGBMXT_BAG_L1	64.7 ± 0.8	78.6 ± 1.4	47.6 ± 1.5	0.277 ± 0.017
LightGBMXT_BAG_L2	66.9 ± 0.7	81.2 ± 2.1	49.3 ± 2.2	0.325 ± 0.015
LightGBM_BAG_L1	64.7 ± 0.8	78.6 ± 1.4	47.6 ± 1.5	0.277 ± 0.017
LightGBM_BAG_L2	67.2 ± 0.6	81.7 ± 1.0	49.3 ± 1.6	0.331 ± 0.012
NeuralNetFastAI_BAG_L1	67.2 ± 0.6	80.9 ± 1.2	50.3 ± 1.5	0.331 ± 0.013
NeuralNetFastAI_BAG_L2	67.3 ± 0.4	80.7 ± 2.0	50.7 ± 2.2	0.332 ± 0.009
NeuralNetTorch_BAG_L1	66.0 ± 0.5	79.4 ± 2.7	49.5 ± 3.0	0.305 ± 0.010
NeuralNetTorch_BAG_L2	66.7 ± 0.3	78.9 ± 2.2	51.6 ± 3.2	0.319 ± 0.007
RandomForestEntr_BAG_L1	67.9 ± 0.3	82.3 ± 1.3	50.1 ± 2.0	0.345 ± 0.005
RandomForestEntr_BAG_L2	67.3 ± 0.1	80.9 ± 1.3	50.6 ± 1.7	0.332 ± 0.003
RandomForestGini_BAG_L1	67.8 ± 0.4	82.5 ± 1.0	49.7 ± 1.7	0.343 ± 0.008
RandomForestGini_BAG_L2	67.3 ± 0.4	81.2 ± 1.0	50.1 ± 1.9	0.332 ± 0.007
WeightedEnsemble_L2	66.8 ± 1.3	81.4 ± 1.8	48.9 ± 2.4	0.322 ± 0.028
WeightedEnsemble_L3	67.3 ± 0.7	80.9 ± 1.1	50.5 ± 2.3	0.332 ± 0.014
XGBoost_BAG_L1	63.7 ± 1.0	75.6 ± 1.6	49.0 ± 1.6	0.256 ± 0.021
XGBoost_BAG_L2	67.4 ± 0.4	81.2 ± 1.0	50.3 ± 1.8	0.333 ± 0.009

Table 25: Test performance for transfer learning of models developed on developmental toxicity and used to predict on reproductive toxicity data

Model name	ACC (%)	SE (%)	SP (%)	MCC
CatBoost_BAG_L1	77.2 ± 1.7	84.4 ± 0.9	69.2 ± 4.4	0.544 ± 0.032
CatBoost_BAG_L2	81.4 ± 1.4	84.3 ± 0.8	78.2 ± 2.5	0.627 ± 0.028
ExtraTreesEntr_BAG_L1	80.3 ± 1.3	88.0 ± 1.2	71.7 ± 2.4	0.608 ± 0.025
ExtraTreesEntr_BAG_L2	80.2 ± 1.6	85.5 ± 0.6	74.2 ± 3.7	0.603 ± 0.030
ExtraTreesGini_BAG_L1	79.9 ± 0.8	88.0 ± 1.2	70.9 ± 2.0	0.601 ± 0.015
ExtraTreesGini_BAG_L2	80.1 ± 1.7	85.6 ± 0.4	74.0 ± 3.7	0.602 ± 0.033
LightGBMLarge_BAG_L1	79.4 ± 1.3	86.9 ± 0.7	71.1 ± 2.2	0.590 ± 0.025
LightGBMLarge_BAG_L2	78.7 ± 0.9	84.7 ± 0.5	72.0 ± 2.3	0.573 ± 0.018
LightGBMXT_BAG_L1	76.8 ± 1.5	82.6 ± 1.4	70.4 ± 2.5	0.536 ± 0.029
LightGBMXT_BAG_L2	80.0 ± 1.6	85.8 ± 0.9	73.4 ± 3.8	0.599 ± 0.030
LightGBM_BAG_L1	76.8 ± 1.5	82.6 ± 1.4	70.4 ± 2.5	0.536 ± 0.029
LightGBM_BAG_L2	79.0 ± 0.9	85.1 ± 1.0	72.1 ± 2.6	0.580 ± 0.017
NeuralNetFastAI_BAG_L1	77.0 ± 1.1	85.3 ± 1.1	67.9 ± 1.9	0.542 ± 0.022
NeuralNetFastAI_BAG_L2	76.5 ± 2.1	84.7 ± 1.1	67.3 ± 3.4	0.530 ± 0.043
NeuralNetTorch_BAG_L1	77.1 ± 1.2	85.9 ± 0.5	67.3 ± 2.3	0.544 ± 0.022
NeuralNetTorch_BAG_L2	79.1 ± 2.4	83.6 ± 1.2	74.0 ± 4.1	0.580 ± 0.049
RandomForestEntr_BAG_L1	80.4 ± 1.2	87.8 ± 0.7	72.0 ± 2.5	0.609 ± 0.024
RandomForestEntr_BAG_L2	80.4 ± 1.3	84.5 ± 0.4	75.7 ± 2.7	0.606 ± 0.026
RandomForestGini_BAG_L1	80.1 ± 1.2	88.0 ± 1.1	71.2 ± 2.4	0.604 ± 0.024
RandomForestGini_BAG_L2	80.4 ± 1.3	84.6 ± 0.6	75.6 ± 3.3	0.606 ± 0.026
WeightedEnsemble_L2	80.0 ± 1.0	87.0 ± 2.2	72.1 ± 2.1	0.600 ± 0.022
WeightedEnsemble_L3	79.9 ± 1.6	85.0 ± 1.1	74.2 ± 4.5	0.598 ± 0.032
XGBoost_BAG_L1	76.8 ± 1.2	83.0 ± 1.0	69.9 ± 2.8	0.536 ± 0.024
XGBoost_BAG_L2	79.9 ± 1.1	84.6 ± 1.0	74.5 ± 2.0	0.596 ± 0.021

Table 26: Results of similarity method (Chapter 3) between developmental toxicity and reproductive toxicity datasets

Tanimoto similarity threshold used	"Similarity" of reproductive toxicity data with developmental toxicity data (%)	"Similarity" of developmental toxicity data with reproductive toxicity data (%)	Average "Similarity" between both datasets (%)
0.2	93.6	98.4	95.6
0.3	70.4	88.7	79.5

In addition, the similarity between the datasets were analysed using the similarity method in Chapter 3. The results are shown in Table 26. Based on the results, both similarity thresholds indicate that the models trained on one toxicity endpoint are applicable to the other toxicity endpoint as the average similarity between datasets (S) is 70% or higher. This also supports the earlier ML results in Table 24 and Table 25. It is interesting to note that the ML results obtained here are comparable to those summarised in Table 16, where the models were trained from scratch for each endpoint. This could mean that the poor performance of ML models for developmental toxicity is due to compounds not commonly seen for reproductive toxicity *i.e.* they probably target different MIEs and AOPs. A quick check of the datasets for both toxicity endpoints shows that about 600 compounds are common between both endpoints, which corresponds to about 27% and 38% for developmental and reproductive toxicity respectively. This is unlikely to significantly influence the current ML results or be the reason for the low performance of the ML models for developmental toxicity.

4.3.11 Transfer learning between developmental toxicity/reproductive toxicity models and human targets

In order to better understand the mechanisms of action involved in DART, transfer learning between the models for developmental toxicity/reproductive toxicity was carried out with the 79 sets of human target data in the work by Allen et al which was also used in Chapter 3.[358] While transfer learning is useful for endpoints with a scarcity of data, data scarcity is not a problem for the receptors investigated in this section. Therefore, the main goal of this section is to identify possible mechanisms of action involved in DART. With the knowledge that when the transfer learning of developmental toxicity/reproductive toxicity models on the human

target data results in good model performance, it represents that the molecules in the two datasets are similar enough. This is because the patterns learned by the developmental toxicity/reproductive toxicity models can also be applied to the human target data. The transferability of these patterns suggests that the binding to a particular receptor could be one of the mechanisms leading to developmental toxicity/reproductive toxicity.

The models for all 5 runs per toxicity type (developmental toxicity/reproductive toxicity) were applied to each of the test datasets for the 79 targets and the results for the best-performing model for each target are summarised. Table 27 and Table 28 show the results obtained for the models of developmental toxicity.

Table 27: Results of best-performing models obtained by applying transfer learning of developmental toxicity models to human target test datasets

No.	Target	ACC (%)	SE (%)	SP (%)	MCC
1	AChE	62.6 ± 1.9	83.7 ± 4.6	36.7 ± 2.0	0.235 ± 0.044
2	ADORA2A	75.9 ± 1.7	86.6 ± 2.9	54.7 ± 1.7	0.440 ± 0.035
3	ADRA2A	68.3 ± 1.1	84.9 ± 1.2	55.1 ± 2.4	0.411 ± 0.017
4	ADRB1	70.9 ± 2.6	90.2 ± 6.2	49.8 ± 1.7	0.444 ± 0.067
5	ADRB2	63.2 ± 0.3	86.8 ± 3.9	41.0 ± 3.7	0.313 ± 0.015
6	AGTR1	68.6 ± 1.6	91.2 ± 4.5	53.3 ± 2.6	0.458 ± 0.043
7	AKT1	81.4 ± 1.2	93.8 ± 2.4	54.5 ± 2.5	0.548 ± 0.029
8	AR	54.2 ± 0.9	79.3 ± 3.4	45.2 ± 1.5	0.222 ± 0.026
9	AVPR1A	66.2 ± 2.6	84.8 ± 6.4	55.1 ± 1.9	0.395 ± 0.069
10	BACE1	78.4 ± 1.2	91.4 ± 2.2	46.7 ± 1.6	0.437 ± 0.028
11	BCHE	59.3 ± 1.9	78.7 ± 5.7	46.1 ± 1.5	0.255 ± 0.054
12	CASP1	46.9 ± 0.4	77.7 ± 10.0	32.5 ± 5.2	0.108 ± 0.053
13	CASP3	60.9 ± 1.7	79.3 ± 10.9	49.4 ± 4.1	0.292 ± 0.079
14	CASP8	56.7 ± 2.3	69.8 ± 12.7	52.9 ± 1.6	0.191 ± 0.100
15	CHRM1	71.3 ± 2.8	83.0 ± 4.5	51.0 ± 1.8	0.362 ± 0.056
16	CHRM2	65.4 ± 1.4	79.2 ± 2.1	54.6 ± 1.8	0.343 ± 0.028
17	CHRM3	67.3 ± 1.7	82.3 ± 3.7	48.4 ± 1.6	0.330 ± 0.036
18	CHRM5	67.5 ± 0.9	89.8 ± 2.5	53.9 ± 0.4	0.438 ± 0.026
19	CHUK	64.4 ± 1.6	92.8 ± 3.6	54.9 ± 2.3	0.418 ± 0.031

20	CSF1R	72.5 ± 3.1	85.0 ± 7.3	56.5 ± 3.1	0.441 ± 0.064
21	CSNK1D	65.4 ± 1.8	81.6 ± 5.4	53.4 ± 1.9	0.357 ± 0.049
22	DRD1	69.9 ± 0.6	90.6 ± 1.0	56.8 ± 1.1	0.475 ± 0.012
23	DRD2	88.5 ± 1.7	96.3 ± 2.0	48.9 ± 0.8	0.538 ± 0.059
24	EDNRA	63.2 ± 1.6	76.7 ± 2.9	48.4 ± 2.3	0.263 ± 0.034
25	EDNRB	62.9 ± 1.8	81.9 ± 6.2	50.2 ± 2.2	0.327 ± 0.056
26	ELANE	69.9 ± 1.6	80.9 ± 2.6	53.3 ± 3.0	0.357 ± 0.036
27	EPHA2	69.4 ± 1.3	93.2 ± 3.8	57.1 ± 1.6	0.487 ± 0.036
28	FGFR1	76.8 ± 1.5	90.6 ± 3.1	51.6 ± 2.6	0.472 ± 0.034
29	FKBP1A	61.7 ± 1.4	96.8 ± 1.8	48.5 ± 2.3	0.421 ± 0.014
30	FLT1	66.1 ± 2.7	81.1 ± 7.3	56.8 ± 1.7	0.374 ± 0.072
31	FLT4	70.7 ± 2.0	87.6 ± 6.0	58.3 ± 2.2	0.467 ± 0.056
32	FYN	58.8 ± 0.9	83.2 ± 4.6	49.8 ± 0.8	0.298 ± 0.038
33	GSK3B	72.5 ± 3.2	85.3 ± 5.7	48.4 ± 2.3	0.368 ± 0.059
34	HDAC3	59.0 ± 5.1	62.2 ± 12.8	55.8 ± 2.7	0.184 ± 0.108
35	HRH1	71.3 ± 0.7	93.0 ± 2.4	47.5 ± 1.7	0.460 ± 0.022
36	HTR2A	84.8 ± 1.7	92.1 ± 2.2	57.9 ± 2.5	0.529 ± 0.042
37	HTR3A	60.5 ± 1.1	68.1 ± 2.9	57.7 ± 2.5	0.230 ± 0.010
38	IGF1R	78.3 ± 2.6	90.8 ± 4.1	54.0 ± 1.8	0.497 ± 0.058
39	INSR	70.0 ± 2.9	82.0 ± 6.9	59.5 ± 1.9	0.424 ± 0.068
40	KCNH2	65.7 ± 1.0	89.8 ± 2.2	28.3 ± 2.1	0.235 ± 0.027
41	KDR	75.9 ± 3.8	81.3 ± 4.7	47.8 ± 2.0	0.255 ± 0.046
42	LCK	72.1 ± 5.2	79.0 ± 7.4	46.1 ± 5.4	0.236 ± 0.062
43	LTB4R	56.2 ± 1.7	59.4 ± 11.2	55.0 ± 2.1	0.128 ± 0.083
44	LYN	60.7 ± 2.8	79.5 ± 8.5	52.7 ± 2.1	0.298 ± 0.079
45	MAPK1	55.4 ± 0.8	70.4 ± 5.3	47.2 ± 3.9	0.172 ± 0.018
46	MAPK9	69.0 ± 2.8	81.1 ± 5.8	54.8 ± 0.9	0.376 ± 0.059
47	MAPKAPK2	70.8 ± 1.0	90.7 ± 1.0	55.9 ± 1.8	0.480 ± 0.017
48	MET	72.9 ± 4.5	83.4 ± 7.6	49.7 ± 2.5	0.351 ± 0.072
49	MMP13	71.4 ± 3.0	79.7 ± 5.1	56.0 ± 1.5	0.365 ± 0.054
50	MMP2	69.6 ± 1.9	83.1 ± 3.0	45.1 ± 2.5	0.306 ± 0.040
51	MMP3	74.5 ± 2.7	86.7 ± 4.8	53.4 ± 3.1	0.434 ± 0.059

52	MMP9	68.0 ± 1.3	77.8 ± 2.0	54.5 ± 1.5	0.333 ± 0.027
53	NEK2	61.3 ± 1.1	81.1 ± 9.2	55.9 ± 1.7	0.303 ± 0.065
54	NR3C1	55.8 ± 1.1	79.1 ± 3.8	45.8 ± 1.3	0.235 ± 0.034
55	OPRD1	73.9 ± 8.2	83.3 ± 12.1	50.9 ± 1.4	0.370 ± 0.143
56	OPRM1	73.1 ± 2.7	87.3 ± 4.7	50.0 ± 1.3	0.411 ± 0.060
57	P2RY1	59.9 ± 2.4	69.6 ± 5.8	55.4 ± 2.6	0.233 ± 0.057
58	PAK4	64.7 ± 1.9	88.9 ± 5.4	57.0 ± 1.7	0.394 ± 0.050
59	PDE4A	64.0 ± 3.4	75.4 ± 9.2	55.6 ± 2.1	0.312 ± 0.087
60	PDE5A	64.5 ± 2.6	74.8 ± 4.7	51.9 ± 3.1	0.276 ± 0.054
61	PIK3CA	72.3 ± 4.9	81.1 ± 7.5	53.1 ± 1.8	0.353 ± 0.080
62	PPARG	53.3 ± 2.5	70.4 ± 6.2	42.9 ± 1.5	0.133 ± 0.065
63	PTPN1	56.4 ± 1.4	75.2 ± 7.5	44.9 ± 3.9	0.204 ± 0.050
64	PTPN11	58.7 ± 0.6	80.9 ± 5.4	52.1 ± 1.4	0.279 ± 0.037
65	PTPN2	57.6 ± 1.7	65.4 ± 6.0	55.5 ± 1.1	0.172 ± 0.052
66	RAF1	72.7 ± 3.4	82.9 ± 6.9	60.1 ± 2.3	0.447 ± 0.074
67	RARA	61.5 ± 1.6	71.1 ± 2.3	60.2 ± 1.9	0.202 ± 0.015
68	RARB	58.4 ± 1.8	80.0 ± 4.8	56.5 ± 2.2	0.198 ± 0.021
69	ROCK1	72.9 ± 1.3	89.1 ± 3.8	54.4 ± 2.3	0.470 ± 0.032
70	RPS6KA5	62.0 ± 1.5	83.2 ± 8.3	57.4 ± 2.8	0.311 ± 0.049
71	SIRT2	56.2 ± 1.6	73.7 ± 9.9	51.1 ± 2.9	0.208 ± 0.067
72	SIRT3	61.4 ± 2.2	79.5 ± 10.1	58.4 ± 3.3	0.264 ± 0.056
73	SLC6A2	76.8 ± 1.2	91.4 ± 2.5	56.3 ± 2.5	0.522 ± 0.029
74	SLC6A3	76.3 ± 0.4	90.4 ± 2.1	57.2 ± 2.4	0.514 ± 0.012
75	SLC6A4	85.3 ± 1.6	93.8 ± 2.2	55.9 ± 2.3	0.549 ± 0.043
76	SRC	70.5 ± 4.1	84.1 ± 7.1	48.8 ± 3.0	0.359 ± 0.081
77	TACR2	66.0 ± 2.2	86.9 ± 5.5	56.2 ± 1.9	0.407 ± 0.057
78	TBXA2R	60.9 ± 1.7	75.9 ± 5.5	53.9 ± 0.5	0.280 ± 0.052
79	TEK	65.0 ± 1.2	79.3 ± 5.3	54.5 ± 2.1	0.341 ± 0.038

Table 28: Comparison of the best-performing models obtained by applying transfer learning of developmental toxicity models to human target test datasets with the best-performing models for developmental toxicity

No.	Target	Best-performing model obtained by applying transfer learning	Is this a different model as compared to the best-performing model for developmental toxicity
1	AChE	RandomForestGini_BAG_L1	Yes
2	ADORA2A	ExtraTreesEntr_BAG_L1	Yes
3	ADRA2A	ExtraTreesEntr_BAG_L1	Yes
4	ADRB1	ExtraTreesEntr_BAG_L1	Yes
5	ADRB2	RandomForestEntr_BAG_L1	Yes
6	AGTR1	ExtraTreesEntr_BAG_L1	Yes
7	AKT1	RandomForestGini_BAG_L1	Yes
8	AR	ExtraTreesGini_BAG_L1	Yes
9	AVPR1A	RandomForestEntr_BAG_L1	Yes
10	BACE1	ExtraTreesEntr_BAG_L1	Yes
11	BCHE	RandomForestGini_BAG_L1	Yes
12	CASP1	ExtraTreesEntr_BAG_L1	Yes
13	CASP3	RandomForestEntr_BAG_L1	Yes
14	CASP8	ExtraTreesGini_BAG_L1	Yes
15	CHRM1	ExtraTreesGini_BAG_L1	Yes
16	CHRM2	RandomForestGini_BAG_L1	Yes
17	CHRM3	RandomForestGini_BAG_L1	Yes
18	CHRM5	ExtraTreesGini_BAG_L1	Yes
19	CHUK	RandomForestGini_BAG_L1	Yes
20	CSF1R	ExtraTreesEntr_BAG_L1	Yes
21	CSNK1D	RandomForestGini_BAG_L1	Yes
22	DRD1	RandomForestGini_BAG_L1	Yes
23	DRD2	ExtraTreesGini_BAG_L1	Yes
24	EDNRA	RandomForestGini_BAG_L1	Yes
25	EDNRB	RandomForestEntr_BAG_L1	Yes
26	ELANE	ExtraTreesEntr_BAG_L1	Yes

27	EPHA2	RandomForestGini_BAG_L1	Yes
28	FGFR1	RandomForestGini_BAG_L1	Yes
29	FKBP1A	RandomForestEntr_BAG_L1	Yes
30	FLT1	ExtraTreesEntr_BAG_L1	Yes
31	FLT4	RandomForestGini_BAG_L1	Yes
32	FYN	ExtraTreesEntr_BAG_L1	Yes
33	GSK3B	ExtraTreesGini_BAG_L1	Yes
34	HDAC3	LightGBMLarge_BAG_L1	Yes
35	HRH1	RandomForestGini_BAG_L1	Yes
36	HTR2A	ExtraTreesEntr_BAG_L1	Yes
37	HTR3A	ExtraTreesEntr_BAG_L1	Yes
38	IGF1R	ExtraTreesEntr_BAG_L1	Yes
39	INSR	ExtraTreesEntr_BAG_L1	Yes
40	KCNH2	RandomForestGini_BAG_L1	Yes
41	KDR	ExtraTreesEntr_BAG_L1	Yes
42	LCK	RandomForestEntr_BAG_L1	Yes
43	LTB4R	RandomForestEntr_BAG_L1	Yes
44	LYN	ExtraTreesEntr_BAG_L1	Yes
45	MAPK1	NeuralNetTorch_BAG_L2	Yes
46	MAPK9	ExtraTreesEntr_BAG_L1	Yes
47	MAPKAPK2	ExtraTreesGini_BAG_L1	Yes
48	MET	RandomForestGini_BAG_L1	Yes
49	MMP13	RandomForestGini_BAG_L1	Yes
50	MMP2	ExtraTreesEntr_BAG_L1	Yes
51	MMP3	ExtraTreesEntr_BAG_L1	Yes
52	MMP9	ExtraTreesEntr_BAG_L1	Yes
53	NEK2	RandomForestEntr_BAG_L1	Yes
54	NR3C1	RandomForestGini_BAG_L1	Yes
55	OPRD1	ExtraTreesGini_BAG_L2	Yes
56	OPRM1	RandomForestEntr_BAG_L1	Yes
57	P2RY1	ExtraTreesEntr_BAG_L1	Yes
58	PAK4	RandomForestGini_BAG_L1	Yes

59	PDE4A	ExtraTreesGini_BAG_L1	Yes
60	PDE5A	RandomForestEntr_BAG_L1	Yes
61	PIK3CA	ExtraTreesGini_BAG_L1	Yes
62	PPARG	ExtraTreesEntr_BAG_L1	Yes
63	PTPN1	ExtraTreesEntr_BAG_L1	Yes
64	PTPN11	ExtraTreesEntr_BAG_L1	Yes
65	PTPN2	NeuralNetTorch_BAG_L1	Yes
66	RAF1	RandomForestEntr_BAG_L1	Yes
67	RARA	ExtraTreesGini_BAG_L2	Yes
68	RARB	ExtraTreesEntr_BAG_L1	Yes
69	ROCK1	ExtraTreesEntr_BAG_L1	Yes
70	RPS6KA5	ExtraTreesGini_BAG_L1	Yes
71	SIRT2	ExtraTreesEntr_BAG_L1	Yes
72	SIRT3	XGBoost_BAG_L1	Yes
73	SLC6A2	ExtraTreesGini_BAG_L1	Yes
74	SLC6A3	RandomForestGini_BAG_L1	Yes
75	SLC6A4	ExtraTreesEntr_BAG_L1	Yes
76	SRC	ExtraTreesEntr_BAG_L1	Yes
77	TACR2	RandomForestEntr_BAG_L1	Yes
78	TBXA2R	ExtraTreesEntr_BAG_L1	Yes
79	TEK	RandomForestGini_BAG_L1	Yes

Generally, it is observed that the values for sensitivity are high while the values for specificity are low. This means that there are few false negatives while there are many false positives. For predictive toxicology, high SE is important so as to not miss any toxic compound for the hazard assessment. When considering the accuracies of the predictions, arbitrary cutoffs have been used where accuracies with $\leq 60\%$ are tagged as "unrelated", $60\% < Acc \leq 70\%$ are treated as "lowly related", $70\% < Acc \leq 80\%$ as "moderately related", and $> 80\%$ as "highly related". The cutoffs facilitate the analysis of when the model performance using transfer learning is regarded as "good". For developmental toxicity, 16 targets have the "unrelated" tag, 35 targets have the "lowly related" tag, 24 targets have the "moderately related" tag, and 4 targets have the "highly related" tag. In the case of developmental toxicity, the results of transfer learning predict high relations with the human targets AKT1, DRD2, HTR2A, and SLC6A4.

AKT1 is related to the adrenal IGF1 signaling pathway which plays a vital role in the body's normal development of multiple organs.[434] Chemicals aimed at AKT1 were also screened for developmental toxicity.[435] DRD2 has been found to be involved in developmental toxicity to the cardiovascular and nervous system of zebrafish larvae.[436] DRD2 has also been shown to be related to prenatal developmental toxicity.[365] HTR2A is related to developmental neurotoxicity in zebrafish larvae.[437] HTR2A is also an important regulator of fetal brain development and adult cognitive function.[438] Changes in the expression of SLC6A4 which is a 5-HT transporter could result in altered neurodevelopment during the mid-gestational window or adverse neonatal outcomes in infants.[439, 440]

The literature thus supports the results of transfer learning and thus confirms that patterns learned by the developmental toxicity models can also be applied to the human target data. The information learned can be applied to the development of adverse outcome pathways for developmental toxicity. Also, by investigating the results of transfer learning on the different human targets, more targeted *in vitro* screening/testing can be conducted in order to better understand the mechanisms behind developmental toxicity.

Also, none of the best-performing models for the transfer learning process are the same as those trained on the developmental toxicity data as summarised in Table 28. For transfer learning with the models of developmental toxicity on human target data, the random forest models are generally the best-performing model which suggests that these models are perhaps more suitable for transfer learning as compared to other ML models.

A similar investigation was carried out for reproductive toxicity with the results shown in Table 29 and Table 30. Once again, the values for sensitivity are high while the values for specificity are low. As compared to developmental toxicity, the values are generally higher for sensitivity and lower for specificity. Using the same tags for the accuracies, it was determined that 31 targets have the “unrelated” tag, 24 targets have the “lowly related” tag, 16 targets have the “moderately related” tag, and 8 targets have the “highly related” tag. The human targets that are predicted to be highly related to reproductive toxicity are ADORA2A, BACE1, DRD2, HTR2A, KDR, LCK, PIK3CA, and SLC6A4.

Table 29: Results of best-performing models obtained by applying transfer learning of reproductive toxicity models to human target test datasets

No.	Target	ACC (%)	SE (%)	SP (%)	MCC
1	AChE	62.9 ± 0.2	93.7 ± 0.9	25.0 ± 1.2	0.264 ± 0.007
2	ADORA2A	81.2 ± 1.1	98.3 ± 0.6	47.4 ± 4.1	0.574 ± 0.024
3	ADRA2A	60.0 ± 1.0	94.2 ± 0.7	32.6 ± 1.6	0.328 ± 0.019
4	ADRB1	66.4 ± 0.7	98.2 ± 0.6	31.7 ± 1.7	0.407 ± 0.012
5	ADRB2	57.3 ± 0.7	95.6 ± 0.5	21.2 ± 1.3	0.249 ± 0.014
6	AGTR1	59.9 ± 2.0	96.6 ± 0.5	35.0 ± 3.6	0.373 ± 0.025
7	AKT1	77.8 ± 0.3	99.4 ± 0.2	31.0 ± 1.0	0.467 ± 0.009
8	AR	55.9 ± 2.2	86.8 ± 1.4	44.9 ± 3.5	0.29 ± 0.015
9	AVPR1A	59.7 ± 1.9	98.7 ± 0.8	36.3 ± 2.9	0.401 ± 0.028
10	BACE1	80.4 ± 0.5	98.9 ± 0.1	35.2 ± 1.9	0.496 ± 0.016
11	BCHE	62.0 ± 2.2	93.5 ± 1.3	40.6 ± 4.3	0.378 ± 0.026
12	CASP1	38.5 ± 0.6	98.3 ± 1.1	10.7 ± 0.8	0.156 ± 0.021
13	CASP3	50.9 ± 0.3	95.8 ± 1.1	22.9 ± 1.0	0.25 ± 0.009
14	CASP8	51.9 ± 1.7	96.0 ± 1.1	38.9 ± 2.4	0.317 ± 0.013
15	CHRM1	71.7 ± 0.5	95.1 ± 0.2	31.0 ± 1.1	0.359 ± 0.013
16	CHRM2	67.7 ± 1.2	93.5 ± 0.8	47.5 ± 2.2	0.446 ± 0.02
17	CHRM3	67.1 ± 1.0	95.4 ± 0.9	31.4 ± 1.8	0.359 ± 0.026
18	CHRM5	61.3 ± 1.2	97.5 ± 0.5	39.3 ± 2.0	0.41 ± 0.013
19	CHUK	51.5 ± 2.9	98.9 ± 0.0	35.7 ± 3.9	0.338 ± 0.026
20	CSF1R	71.5 ± 1.6	98.9 ± 0.3	36.5 ± 4.0	0.472 ± 0.028
21	CSNK1D	61.6 ± 1.3	98.2 ± 0.3	34.4 ± 2.2	0.399 ± 0.021
22	DRD1	66.5 ± 1.9	93.7 ± 1.0	49.1 ± 3.7	0.446 ± 0.02
23	DRD2	88.1 ± 0.2	99.2 ± 0.2	31.4 ± 1.4	0.484 ± 0.012
24	EDNRA	66.3 ± 0.8	98.6 ± 0.4	30.6 ± 1.8	0.406 ± 0.015
25	EDNRB	59.4 ± 1.0	99.6 ± 0.5	32.4 ± 1.8	0.396 ± 0.014
26	ELANE	69.4 ± 0.5	93.0 ± 1.4	33.6 ± 3.3	0.343 ± 0.009
27	EPHA2	57.1 ± 1.4	98.3 ± 0.6	35.8 ± 2.2	0.377 ± 0.017
28	FGFR1	76.6 ± 0.8	99.6 ± 0.2	34.5 ± 2.5	0.493 ± 0.019
29	FKBP1A	52.2 ± 1.9	98.8 ± 0.4	34.5 ± 2.7	0.342 ± 0.016

30	FLT1	66.8 ± 0.9	99.3 ± 0.1	46.7 ± 1.5	0.492 ± 0.012
31	FLT4	63.5 ± 2.2	97.7 ± 1.0	38.4 ± 4.5	0.424 ± 0.024
32	FYN	53.0 ± 1.6	97.6 ± 0.6	36.7 ± 2.2	0.34 ± 0.016
33	GSK3B	75.0 ± 0.2	98.5 ± 0.4	30.5 ± 1.0	0.433 ± 0.007
34	HDAC3	62.3 ± 1.4	91.1 ± 2.0	33.1 ± 4.3	0.298 ± 0.023
35	HRH1	68.1 ± 0.8	98.5 ± 0.2	34.5 ± 1.6	0.437 ± 0.014
36	HTR2A	85.2 ± 0.3	99.0 ± 0.4	35.0 ± 2.3	0.505 ± 0.013
37	HTR3A	51.3 ± 1.4	92.7 ± 1.3	35.7 ± 2.0	0.282 ± 0.018
38	IGF1R	77.3 ± 0.8	99.3 ± 0.3	34.6 ± 2.8	0.49 ± 0.02
39	INSR	66.0 ± 0.9	98.3 ± 1.7	37.8 ± 2.7	0.444 ± 0.016
40	KCNH2	65.6 ± 0.2	98.3 ± 0.3	15.1 ± 0.7	0.256 ± 0.007
41	KDR	87.8 ± 0.3	98.7 ± 0.1	30.9 ± 2.1	0.455 ± 0.019
42	LCK	80.7 ± 0.6	98.3 ± 0.5	14.6 ± 2.8	0.254 ± 0.039
43	LTB4R	52.4 ± 0.9	95.5 ± 1.4	36.1 ± 1.6	0.316 ± 0.008
44	LYN	53.3 ± 1.7	97.5 ± 0.4	34.3 ± 2.3	0.337 ± 0.019
45	MAPK1	44.3 ± 1.3	91.6 ± 2.2	18.2 ± 3.2	0.133 ± 0.006
46	MAPK9	67.9 ± 1.1	99.1 ± 0.7	31.3 ± 3.2	0.426 ± 0.015
47	MAPKAPK2	60.4 ± 0.7	97.9 ± 0.8	32.4 ± 1.4	0.38 ± 0.01
48	MET	79.3 ± 0.5	99.0 ± 0.5	35.4 ± 2.7	0.496 ± 0.01
49	MMP13	72.0 ± 0.9	93.6 ± 1.2	31.7 ± 1.6	0.337 ± 0.027
50	MMP2	67.3 ± 1.5	91.3 ± 1.7	24.0 ± 1.8	0.211 ± 0.042
51	MMP3	70.8 ± 1.6	91.8 ± 1.9	34.7 ± 1.4	0.334 ± 0.044
52	MMP9	66.7 ± 1.2	86.4 ± 1.5	39.4 ± 0.9	0.297 ± 0.028
53	NEK2	47.7 ± 1.6	98.6 ± 1.0	34.0 ± 2.0	0.301 ± 0.017
54	NR3C1	58.6 ± 1.0	89.3 ± 1.0	45.3 ± 1.7	0.333 ± 0.01
55	OPRD1	79.2 ± 0.9	99.1 ± 0.2	30.7 ± 2.6	0.461 ± 0.029
56	OPRM1	77.3 ± 0.6	98.0 ± 0.3	43.6 ± 1.8	0.527 ± 0.012
57	P2RY1	57.4 ± 1.9	98.0 ± 0.4	38.7 ± 2.6	0.384 ± 0.023
58	PAK4	52.1 ± 1.5	100.0 ± 0.0	37.1 ± 1.9	0.352 ± 0.013
59	PDE4A	61.7 ± 1.1	98.7 ± 0.3	34.3 ± 2.0	0.406 ± 0.014
60	PDE5A	69.0 ± 0.4	98.8 ± 0.4	32.6 ± 1.2	0.434 ± 0.006
61	PIK3CA	82.0 ± 0.6	99.8 ± 0.1	43.0 ± 2.0	0.58 ± 0.014

62	PPARG	58.8 ± 1.1	85.0 ± 0.8	42.7 ± 2.0	0.288 ± 0.013
63	PTPN1	49.7 ± 0.4	95.4 ± 0.8	21.6 ± 0.8	0.23 ± 0.01
64	PTPN11	44.4 ± 1.0	90.2 ± 2.9	30.8 ± 2.0	0.202 ± 0.016
65	PTPN2	41.9 ± 1.5	94.6 ± 0.7	27.8 ± 1.9	0.217 ± 0.016
66	RAF1	72.3 ± 0.9	99.3 ± 0.1	38.7 ± 2.0	0.496 ± 0.016
67	RARA	56.3 ± 1.6	90.7 ± 2.6	51.8 ± 2.0	0.273 ± 0.014
68	RARB	55.0 ± 3.0	90.7 ± 1.3	52.0 ± 3.3	0.231 ± 0.014
69	ROCK1	70.8 ± 1.2	99.8 ± 0.3	37.5 ± 2.7	0.489 ± 0.021
70	RPS6KA5	48.5 ± 2.4	97.2 ± 2.4	37.9 ± 3.0	0.289 ± 0.022
71	SIRT2	46.1 ± 1.5	85.8 ± 2.8	34.7 ± 2.5	0.186 ± 0.018
72	SIRT3	41.8 ± 1.6	85.6 ± 8.6	34.7 ± 2.7	0.152 ± 0.053
73	SLC6A2	75.1 ± 0.6	96.2 ± 0.3	45.4 ± 1.5	0.503 ± 0.012
74	SLC6A3	74.7 ± 0.3	94.8 ± 0.6	47.3 ± 1.1	0.495 ± 0.007
75	SLC6A4	83.5 ± 0.5	97.6 ± 0.6	34.2 ± 2.0	0.454 ± 0.02
76	SRC	71.3 ± 0.4	98.8 ± 0.5	27.4 ± 1.1	0.404 ± 0.011
77	TACR2	66.2 ± 2.4	98.8 ± 0.3	50.9 ± 3.5	0.486 ± 0.025
78	TBXA2R	59.7 ± 0.4	91.3 ± 1.9	44.9 ± 1.2	0.358 ± 0.012
79	TEK	62.9 ± 1.1	98.6 ± 0.7	36.8 ± 2.3	0.423 ± 0.013

Table 30: Comparison of the best-performing models obtained by applying transfer learning of reproductive toxicity models to human target test datasets with the best-performing models for reproductive toxicity

No.	Target	Best-performing model obtained by applying transfer learning	Is this a different model as compared to the best-performing model for reproductive toxicity
1	AChE	XGBoost_BAG_L1	Yes
2	ADORA2A	NeuralNetTorch_BAG_L2	Yes
3	ADRA2A	CatBoost_BAG_L1	Yes
4	ADRB1	ExtraTreesEntr_BAG_L1	Yes
5	ADRB2	RandomForestEntr_BAG_L1	No
6	AGTR1	NeuralNetTorch_BAG_L2	Yes
7	AKT1	CatBoost_BAG_L2	Yes

8	AR	NeuralNetTorch_BAG_L2	Yes
9	AVPR1A	CatBoost_BAG_L2	Yes
10	BACE1	CatBoost_BAG_L2	Yes
11	BCHE	NeuralNetTorch_BAG_L2	Yes
12	CASP1	RandomForestEntr_BAG_L1	No
13	CASP3	CatBoost_BAG_L1	Yes
14	CASP8	LightGBMXT_BAG_L1	Yes
15	CHRM1	CatBoost_BAG_L1	Yes
16	CHRM2	CatBoost_BAG_L2	Yes
17	CHRM3	WeightedEnsemble_L2	Yes
18	CHRM5	CatBoost_BAG_L1	Yes
19	CHUK	NeuralNetTorch_BAG_L2	Yes
20	CSF1R	NeuralNetTorch_BAG_L2	Yes
21	CSNK1D	LightGBMXT_BAG_L1	Yes
22	DRD1	NeuralNetTorch_BAG_L2	Yes
23	DRD2	WeightedEnsemble_L2	Yes
24	EDNRA	CatBoost_BAG_L1	Yes
25	EDNRB	XGBoost_BAG_L1	Yes
26	ELANE	NeuralNetTorch_BAG_L2	Yes
27	EPHA2	CatBoost_BAG_L2	Yes
28	FGFR1	CatBoost_BAG_L2	Yes
29	FKBP1A	CatBoost_BAG_L2	Yes
30	FLT1	CatBoost_BAG_L2	Yes
31	FLT4	NeuralNetTorch_BAG_L2	Yes
32	FYN	CatBoost_BAG_L2	Yes
33	GSK3B	RandomForestEntr_BAG_L1	No
34	HDAC3	NeuralNetTorch_BAG_L1	Yes
35	HRH1	ExtraTreesEntr_BAG_L2	Yes
36	HTR2A	ExtraTreesEntr_BAG_L1	Yes
37	HTR3A	CatBoost_BAG_L2	Yes
38	IGF1R	NeuralNetTorch_BAG_L2	Yes
39	INSR	NeuralNetTorch_BAG_L2	Yes

40	KCNH2	RandomForestGini_BAG_L1	Yes
41	KDR	CatBoost_BAG_L2	Yes
42	LCK	ExtraTreesEntr_BAG_L1	Yes
43	LTB4R	ExtraTreesEntr_BAG_L2	Yes
44	LYN	CatBoost_BAG_L2	Yes
45	MAPK1	NeuralNetTorch_BAG_L1	Yes
46	MAPK9	NeuralNetTorch_BAG_L2	Yes
47	MAPKAPK2	XGBoost_BAG_L1	Yes
48	MET	NeuralNetTorch_BAG_L2	Yes
49	MMP13	CatBoost_BAG_L1	Yes
50	MMP2	CatBoost_BAG_L1	Yes
51	MMP3	ExtraTreesGini_BAG_L2	Yes
52	MMP9	CatBoost_BAG_L1	Yes
53	NEK2	CatBoost_BAG_L2	Yes
54	NR3C1	CatBoost_BAG_L2	Yes
55	OPRD1	CatBoost_BAG_L2	Yes
56	OPRM1	CatBoost_BAG_L2	Yes
57	P2RY1	CatBoost_BAG_L2	Yes
58	PAK4	XGBoost_BAG_L1	Yes
59	PDE4A	CatBoost_BAG_L2	Yes
60	PDE5A	CatBoost_BAG_L2	Yes
61	PIK3CA	CatBoost_BAG_L2	Yes
62	PPARG	CatBoost_BAG_L2	Yes
63	PTPN1	ExtraTreesEntr_BAG_L1	Yes
64	PTPN11	ExtraTreesGini_BAG_L1	Yes
65	PTPN2	CatBoost_BAG_L1	Yes
66	RAF1	CatBoost_BAG_L2	Yes
67	RARA	RandomForestEntr_BAG_L2	Yes
68	RARB	NeuralNetTorch_BAG_L2	Yes
69	ROCK1	CatBoost_BAG_L2	Yes
70	RPS6KA5	NeuralNetTorch_BAG_L2	Yes
71	SIRT2	XGBoost_BAG_L1	Yes

72	SIRT3	ExtraTreesGini_BAG_L2	Yes
73	SLC6A2	ExtraTreesGini_BAG_L1	Yes
74	SLC6A3	RandomForestGini_BAG_L1	Yes
75	SLC6A4	ExtraTreesEntr_BAG_L1	Yes
76	SRC	CatBoost_BAG_L2	Yes
77	TACR2	NeuralNetTorch_BAG_L2	Yes
78	TBXA2R	XGBoost_BAG_L1	Yes
79	TEK	CatBoost_BAG_L2	Yes

It was found in the literature that ADORA2A is related to the pathway that causes breast lesions which are part of reproductive toxicity.[441, 442] BACE1 is suggested to be related to an estrogen receptor-dependent mechanism, though several studies in the literature relate BACE1 to developmental neurotoxicity.[443, 444] DRD2 is suggested to have a role for dopaminergic signaling in events such as fertilization, capacitation, and sperm motility.[445] HTR2A is one of the neurotransmission genes which are related to polymorphism that is associated with asthenozoospermia.[446] KDR was reported to play an important role in the maintenance of microcirculation in ruminants testis as a vital local regulatory determinant of testicular functions.[447] LCK is involved in the regulation of spermatogenetic events.[448] PIK3CA could phosphorylate phosphatidylinositol and subsequently be involved in the conduction of the PI3K-AKT signaling pathway.[449] Activation of the PI3K-AKT pathway can promote the proliferation and anti-apoptosis of Sertoli cells (related to sperm production in males).[449] Changes in the expression levels of SLC6A4 could lead to impaired testosterone synthesis, arrest of spermatogenesis, and harmful effects on male fertility.[450]

Once again, the literature supports the results of transfer learning and thus confirms that patterns learned by the reproductive toxicity models can also be applied to the human target data. The information learned can be applied to the development of adverse outcome pathways for reproductive toxicity. By investigating the results of transfer learning on the different human targets, more targeted *in vitro* testing can be conducted in order to better understand the mechanisms behind reproductive toxicity.

From Table 30, only three of the best-performing models for the transfer learning process are the same as those trained on the reproductive toxicity data. For transfer learning with the models of reproductive toxicity on human target data, the CatBoost models which uses gradient boosting on decision trees seem to work better.

Comparing the targets with the "highly related" tags to the best-performing models for the same target taken from Allen et al, it is generally observed that while the sensitivity values of the results of transfer learning outperforms the models by Allen et al, the other performance metrics are much worse.[358] These results are shown in Table 31 and Table 32. For developmental toxicity, the results of transfer learning perform more poorly than the model trained specifically for each target. This situation is similar for reproductive toxicity with the sensitivities being consistently higher than the reported models for all human targets. This makes sense as both developmental toxicity and reproductive toxicity are complex endpoints and do not only follow a single mechanism of action or binding to a human target.

For both developmental toxicity and reproductive toxicity, a good accuracy *i.e.* would indicate that the binding to a human target could be one of the mechanisms leading to the toxicity endpoint. This is because being able to classify the inactives and actives for the human target accurately shows that the model must have learned about the human target from the training dataset of the model. On the other hand, a high sensitivity with a low accuracy would mean that the model could just be predicting actives blindly and is thus unreliable for that particular human target.

When the models for developmental toxicity and reproductive toxicity are applied to larger datasets *i.e.* the training dataset and the combined dataset (test + training), the performance metrics for the best-performing models do not change significantly. This suggests that the datasets are similar in the distribution of compounds. Results for these larger datasets can be found in in the Appendices (Table A17 to Table A20).

Table 31: Comparison of the best-performing models obtained by applying transfer learning of developmental toxicity models to human target test datasets with the best-performing models by Allen et al[358] Only human targets with the "highly related" tags are shown.

Target	This work				Allen et al[358]			
	ACC (%)	SE (%)	SP (%)	MCC	ACC (%)	SE (%)	SP (%)	MCC
AKT1	81.4 ± 1.2	93.8 ± 2.4	54.5 ± 2.5	0.548 ± 0.029	93.9 ± 1.6	96.1 ± 0.9	88.9 ± 4.3	0.854 ± 0.035
DRD2	88.5 ± 1.7	96.3 ± 2.0	48.9 ± 0.8	0.538 ± 0.059	95.6 ± 0.8	98.3 ± 0.8	82.2 ± 4.5	0.838 ± 0.015
HTR2A	84.8 ± 1.7	92.1 ± 2.2	57.9 ± 2.5	0.529 ± 0.042	96.3 ± 0.5	98.5 ± 0.5	88.4 ± 2.4	0.890 ± 0.016
SLC6A4	85.3 ± 1.6	93.8 ± 2.2	55.9 ± 2.3	0.549 ± 0.043	96.6 ± 0.8	98.4 ± 0.5	90.5 ± 3.4	0.901 ± 0.019

Table 32: Comparison of the best-performing models obtained by applying transfer learning of reproductive toxicity models to human target test datasets with the best-performing models by Allen et al**[358]** Only human targets with the "highly related" tags are shown.

Target	This work				Allen et al**[358]**			
	ACC (%)	SE (%)	SP (%)	MCC	ACC (%)	SE (%)	SP (%)	MCC
ADORA2A	81.2 ± 1.1	98.3 ± 0.6	47.4 ± 4.1	0.574 ± 0.024	94.8 ± 0.9	96.7 ± 1.3	91.3 ± 2.0	0.884 ± 0.018
BACE1	80.4 ± 0.5	98.9 ± 0.1	35.2 ± 1.9	0.496 ± 0.016	92.6 ± 0.7	95.8 ± 1.5	84.8 ± 5.1	0.821 ± 0.018
DRD2	88.1 ± 0.2	99.2 ± 0.2	31.4 ± 1.4	0.484 ± 0.012	95.6 ± 0.8	98.3 ± 0.8	82.2 ± 4.5	0.838 ± 0.015
HTR2A	85.2 ± 0.3	99.0 ± 0.4	35.0 ± 2.3	0.505 ± 0.013	96.3 ± 0.5	98.5 ± 0.5	88.4 ± 2.4	0.890 ± 0.016
KDR	87.8 ± 0.3	98.7 ± 0.1	30.9 ± 2.1	0.455 ± 0.019	92.9 ± 0.9	95.7 ± 1.6	79.2 ± 6.9	0.747 ± 0.030
LCK	80.7 ± 0.6	98.3 ± 0.5	14.6 ± 2.8	0.254 ± 0.039	93.8 ± 2.4	95.6 ± 2.3	83.1 ± 9.4	0.805 ± 0.069
PIK3CA	82.0 ± 0.6	99.8 ± 0.1	43.0 ± 2.0	0.58 ± 0.014	96.9 ± 0.6	98.6 ± 0.5	93.1 ± 1.2	0.926 ± 0.013

SLC6A4	83.5 ± 0.5	97.6 ± 0.6	34.2 ± 2.0	0.454 ± 0.02	96.6 ± 0.8	98.4 ± 0.5	90.5 ± 3.4	0.901 ± 0.019

4.3.12 Future outlook

As more data becomes available for DART, the re-training of all models can be considered to cover a larger feature space and applicability domain. To date, the DART database used in the present study is the largest known public database of compounds for the general prediction of the DART endpoint. All of the machine learning model results also indicate that the limit of the performance metrics has been reached given the current DART database. Since the AutoML process is used which has been shown to perform excellently, the improvement of model performance is largely dependent on the quality and quantity of data that can be compiled and obtained from the literature. This emphasises the importance of using good quality data for developing machine learning models and shows that more attention should be placed on ensuring the quality of the data especially when the toxicity endpoints are complex. The training of these models is also expected to be user-friendly and easy to implement with the development of machine learning methods, allowing future improvements to be carried out easily when the need arises.

The use of a group of models that focuses on very specific endpoints is currently expected to be the better choice for DART. This is because the mechanisms and pathways for DART are still not fully explored as evidenced by the lack of AOPs for DART. Thus, ensuring that the toxicity predictions of models are able to be associated with mechanisms of action would be useful as this allows for users to interpret the model results better. While a global model for DART is perhaps easier to use for screening purposes, the use of specific models for developmental toxicity and reproductive separately will probably remain the status quo until more data for DART can be gathered and/or a better mechanistic understanding of pathways leading to DART is achieved.

As mentioned earlier, for both developmental and reproductive toxicity, the literature supports the results of transfer learning. However, one improvement that could be made to the study relates to the receptors with the other tags ie. “unrelated”, “lowly related”, and “moderately related” following the cutoffs explained earlier in section 4.3.11. If literature that supports these tags can be found, it would provide further evidence about the adverse outcome pathways for both developmental toxicity and reproductive toxicity. By proving that a portion of the

"unrelated" tags are indeed unrelated to developmental/reproductive toxicity, more targeted in vitro testing can be carried out based on the results obtained in this study.

4.4 Conclusion

In conclusion, this study has used an automated machine learning (AutoML) process to speed up and optimise the development of in silico models for the binary prediction of DART (toxicants/non-toxicants). This study has also constructed the largest known public database for the general DART effect 3245 compounds (1662 positives, 1583 negatives) that contains data from 12 different sources which includes existing data in the literature such as from datasets used to build *in silico* models. The AutoML process was first benchmarked using literature data to verify its suitability and feasibility before it is applied to the DART database. Subsequently, the AutoML process, which was verified to produce results comparable to the literature, was applied to the new DART database, with a total of 24 machine learning models being trained per run for a total of five runs. The consistent performance of the ML models thus shows that even a complex task such as the prediction of DART toxicants/non-toxicants can be modelled quickly and effectively with AutoML. The predictions made by the models reported in this study can be applied for screening purposes or in Next Generation Risk Assessment (NGRA) frameworks where they could complement other data in the protection of human health.

Transfer learning of these models was also carried out between both toxicity endpoints where with good performance metrics being achieved for some of the models. This shows an overlap between the two toxicity endpoints of developmental toxicity and reproductive toxicity when using this database. The use of transfer learning of the models of developmental toxicity and reproductive toxicity with human target data has also shown possible relationships which can be leveraged for more targeted *in vitro* testing. In particular, human targets AKT1, DRD2, HTR2A, and SLC6A4 have been shown to have links with developmental toxicity in the literature while ADORA2A, BACE1, DRD2, HTR2A, KDR, LCK, PIK3CA, and SLC6A4 have links with reproductive toxicity in the literature. This demonstrates that the use of transfer learning to investigate receptor bindings as possible mechanisms of action leading to DART is viable. This is important as MIE identification is a source of AOP development, especially if the AOP is not yet fully understood or if the AOP is complex. When the results of transfer

learning have a high SE but low SP or ACC, the relationship with the human targets should not be interpreted as they are unreliable.

It is hoped that as more data on DART becomes publicly available, the improvement in performance of *in silico* methods as well as a better understanding of the complex mechanisms leading to DART can support the NGRA.

Chapter 5: Thesis conclusion

Thus far, the investigation of machine learning in predictive toxicology has been carried out. A background of the recent applications and future directions was provided in Chapter 2, while Chapter 3 looks at the use of Tanimoto similarity to calculate average similarity between datasets for human targets. With this method, when the average similarity between datasets (S) or the predicted test accuracy (P) is 70% or higher, it is expected that model performance will be good *i.e.* the model will be transferable to another dataset. This chapter has also shown that the proportion of similar molecules can be used for popular toxicological databases, where about 13 million molecules (PubChem: 13%, ChEMBL: 18%) can be analysed using the Allen et al models for a series of important toxicological effects.[358]

Chapter 4 focuses on DART, where the largest known public database for the general DART effect 3245 compounds (1662 positives, 1583 negatives) was constructed. This database is more up to date with the recent situation given that the study by Wu et al in 2013 which covers a framework to identify structural alerts remains one of the popular choices for DART.[377] Models with about 68% accuracy for developmental toxicity and 80% for reproductive toxicity were trained using the new database. The consistent performance of the ML models for both toxicity endpoints thus shows that even a complex task such as the prediction of DART toxicants/non-toxicants can be modelled quickly and effectively with AutoML. The predictions made by the models reported in this study can be applied for screening purposes or in Next Generation Risk Assessment (NGRA) frameworks where they could complement other results in the protection of human health. The suitability of transfer learning between the two toxicity endpoints has also been investigated with good performance metrics being achieved for some of the models. This shows an overlap between the two toxicity endpoints when using this database. The use of transfer learning of the models of developmental toxicity and reproductive toxicity with human target data has also shown possible relationships between developmental toxicity and reproductive toxicity with human target data which can be leveraged for more targeted *in vitro* testing. In particular, human targets AKT1, DRD2, HTR2A, and SLC6A4 have links with developmental toxicity in the literature while ADORA2A, BACE1, DRD2, HTR2A, KDR, LCK, PIK3CA, and SLC6A4 have links with reproductive toxicity in the literature. The use of transfer learning to investigate receptor bindings as possible mechanisms of action leading to DART has been demonstrated to work in this chapter. These results can be used in MIE/AOP development and targeted *in vitro* screening for DART.

Further improvements to this work can be made. In particular, structural alerts, which are another common tool used in predictive toxicology can be constructed to give an idea of the chemistry behind some of the predictions. It should be noted that generic structural alerts that are generated automatically without proper mechanistic understanding are not as useful as structural alerts that have been manually curated and whose chemistry have been tied to a MIE or AOP. Automatically generated structural alerts can still be useful if they have mechanistic understanding and therefore can be used for identifying the MIE. It should also be noted that manual curation of structural alerts requires expert knowledge which might not be possible in all cases, especially when potential mechanisms are unknown or the AOP is not yet fully understood.

Looking forward, it is expected that machine learning will become an even more powerful tool in predictive toxicology. Recently, the trend has shifted towards explainable machine learning models where understanding a smaller number of predictions is favoured. A better understanding of the pathways/mechanism leading to the endpoints would allow for the generation of more data with mechanistic meaning, which would assist in the development of AOPs.

Chapter 6: References

(1) Wang, M. W. H.; Goodman, J. M.; Allen, T. E. H. Machine Learning in Predictive Toxicology: Recent Applications and Future Directions for Classification Models. *Chemical Research in Toxicology* **2020**, *34* (2), 217–239. DOI: 10.1021/acs.chemrestox.0c00316.

(2) Angermueller, C.; Pärnamaa, T.; Parts, L.; Stegle, O. Deep learning for computational biology. *Mol Syst Biol* **2016**, *12* (7), 878. DOI: 10.15252/msb.20156651 PubMed.

(3) Dana, D.; Gadhiya, S. V.; St. Surin, L. G.; Li, D.; Naaz, F.; Ali, Q.; Paka, L.; Yamin, M. A.; Narayan, M.; Goldberg, I. D.; et al. Deep Learning in Drug Discovery and Medicine; Scratching the Surface. *Molecules* **2018**, *23* (9), 2384. DOI: 10.3390/molecules23092384 PubMed.

(4) Chen, J.; Tang, Y. Y.; Fang, B.; Guo, C. In silico prediction of toxic action mechanisms of phenols for imbalanced data with Random Forest learner. *Journal of Molecular Graphics and Modelling* **2012**, *35*, 21-27. DOI: 10.1016/j.jmgm.2012.01.002.

(5) Merlot, C. Computational toxicology—a tool for early safety evaluation. *Drug Discovery Today* **2010**, *15* (1-2), 16-22. DOI: 10.1016/j.drudis.2009.09.010.

(6) Tang, W.; Chen, J.; Wang, Z.; Xie, H.; Hong, H. Deep learning for predicting toxicity of chemicals: a mini review. *Journal of Environmental Science and Health, Part C* **2018**, *36* (4), 252-271. DOI: 10.1080/10590501.2018.1537563.

(7) Koutsoukas, A.; St. Amand, J.; Mishra, M.; Huan, J. Predictive Toxicology: Modeling Chemical Induced Toxicological Response Combining Circular Fingerprints with Random Forest and Support Vector Machine. *Frontiers in Environmental Science* **2016**, *4*, 11. DOI: 10.3389/fenvs.2016.00011.

(8) Cao, D. S.; Zhao, J. C.; Yang, Y. N.; Zhao, C. X.; Yan, J.; Liu, S.; Hu, Q. N.; Xu, Q. S.; Liang, Y. Z. In silico toxicity prediction by support vector machine and SMILES representation-based string kernel. *SAR and QSAR in Environmental Research* **2012**, *23* (1-2), 141-153. DOI: 10.1080/1062936x.2011.645874.

(9) Kourou, K.; Exarchos, T. P.; Exarchos, K. P.; Karamouzis, M. V.; Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* **2015**, *13*, 8-17. DOI: 10.1016/j.csbj.2014.11.005 PubMed.

(10) Cao, D.-S.; Dong, J.; Wang, N.-N.; Wen, M.; Deng, B.-C.; Zeng, W.-B.; Xu, Q.-S.; Liang, Y.-Z.; Lu, A.-P.; Chen, A. F. In silico toxicity prediction of chemicals from EPA toxicity

database by kernel fusion-based support vector machines. *Chemometrics and Intelligent Laboratory Systems* **2015**, *146*, 494-502. DOI: 10.1016/j.chemolab.2015.07.009.

(11) Wu, K.; Wei, G.-W. Quantitative Toxicity Prediction Using Topology Based Multitask Deep Neural Networks. *Journal of Chemical Information and Modeling* **2018**, *58* (2), 520-531. DOI: 10.1021/acs.jcim.7b00558.

(12) Lavecchia, A. Machine-Learning Approaches in Drug Discovery: Methods and Applications. *Drug Discovery Today* **2015**, *20* (3), 318–331. DOI: 10.1016/j.drudis.2014.10.012.

(13) Mitchell, J. B. O. Machine learning methods in chemoinformatics. *Wiley Interdiscip Rev Comput Mol Sci* **2014**, *4* (5), 468-481. DOI: 10.1002/wcms.1183 PubMed.

(14) Gao, M.; Igata, H.; Takeuchi, A.; Sato, K.; Ikegaya, Y. Machine learning-based prediction of adverse drug effects: An example of seizure-inducing compounds. *Journal of Pharmacological Sciences* **2017**, *133* (2), 70-78. DOI: 10.1016/j.jphs.2017.01.003.

(15) Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discovery Today* **2018**, *23* (8), 1538-1546.

(16) Wu, Y.; Wang, G. Machine Learning Based Toxicity Prediction: From Chemical Structural Description to Transcriptome Analysis. *Int J Mol Sci* **2018**, *19* (8), 2358. DOI: 10.3390/ijms19082358 PubMed.

(17) Idakwo, G.; Luttrell, J.; Chen, M.; Hong, H.; Zhou, Z.; Gong, P.; Zhang, C. A review on machine learning methods for in silico toxicity prediction. *Journal of Environmental Science and Health, Part C* **2018**, *36* (4), 169-191. DOI: 10.1080/10590501.2018.1537118.

(18) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction using Deep Learning. *Frontiers in Environmental Science* **2016**, *3*, 80. DOI: 10.3389/fenvs.2015.00080.

(19) Zhang, L.; Zhang, H.; Ai, H.; Hu, H.; Li, S.; Zhao, J.; Liu, H. Applications of Machine Learning Methods in Drug Toxicity Prediction. *Current Topics in Medicinal Chemistry* **2018**, *18* (12), 987-997. DOI: 10.2174/1568026618666180727152557.

(20) Raies, A. B.; Bajic, V. B. In silico toxicology: comprehensive benchmarking of multi-label classification methods applied to chemical toxicity data. *Wiley Interdiscip Rev Comput Mol Sci* **2018**, *8* (3), e1352. DOI: 10.1002/wcms.1352 PubMed.

(21) Pu, L.; Naderi, M.; Liu, T.; Wu, H.-C.; Mukhopadhyay, S.; Brylinski, M. eToxPred: a machine learning-based approach to estimate the toxicity of drug candidates. *BMC Pharmacol Toxicol* **2019**, *20* (1), 1-15. DOI: 10.1186/s40360-018-0282-6 PubMed.

(22) Lysenko, A.; Sharma, A.; Boroevich, K. A.; Tsunoda, T. An integrative machine learning approach for prediction of toxicity-related drug safety. *Life Sci Alliance* **2018**, *1* (6), e201800098. DOI: 10.26508/lsa.201800098 PubMed.

(23) Wenlock, M. C.; Austin, R. P.; Barton, P.; Davis, A. M.; Leeson, P. D. A Comparison of Physiochemical Property Profiles of Development and Marketed Oral Drugs. *Journal of Medicinal Chemistry* **2003**, *46* (7), 1250-1256. DOI: 10.1021/jm021053p.

(24) Frank, C.; Himmelstein, D. U.; Woolhandler, S.; Bor, D. H.; Wolfe, S. M.; Heymann, O.; Zallman, L.; Lasser, K. E. Era Of Faster FDA Drug Approval Has Also Seen Increased Black-Box Warnings And Market Withdrawals. *Health Affairs* **2014**, *33* (8), 1453-1459. DOI: 10.1377/hlthaff.2014.0122.

(25) Foufelle, F.; Fromenty, B. Role of endoplasmic reticulum stress in drug-induced toxicity. *Pharmacol Res Perspect* **2016**, *4* (1), e00211. DOI: 10.1002/prp2.211 PubMed.

(26) DiMasi, J. Risks in new drug development: Approval success rates for investigational drugs. *Clinical Pharmacology & Therapeutics* **2001**, *69* (5), 297-307. DOI: 10.1067/mcp.2001.115446.

(27) Segall, M. D.; Barber, C. Addressing toxicity risk when designing and selecting compounds in early drug discovery. *Drug Discovery Today* **2014**, *19* (5), 688-693. DOI: 10.1016/j.drudis.2014.01.006.

(28) Fröhlich, E.; Roblegg, E. Models for Oral Uptake of Nanoparticles in Consumer Products. *Toxicology* **2012**, *291* ((1-3)), 10–17.

(29) Mantovani, A.; Maranghi, F.; La Rocca, C.; Tiboni, G. M.; Clementi, M. The Role of Toxicology to Characterize Biomarkers for Agrochemicals with Potential Endocrine Activities. *Reprod. Toxicol.* **2008**, *26* (1), 1–7.

(30) Smith, M.-C.; Madec, S.; Coton, E.; Hymery, N. Natural Co-Occurrence of Mycotoxins in Foods and Feeds and Their in Vitro Combined Toxicological Effects. *Toxins* **2016**, *8* (4), 94.

(31) Rovida, C.; Asakura, S.; Daneshian, M.; Hofman-Huether, H.; Leist, M.; Meunier, L.; Reif, D.; Rossi, A.; Schmutz, M.; Valentin, J. P.; et al. Toxicity testing in the 21st century beyond environmental chemicals. *ALTEX* **2015**, *32* (3), 171-181. DOI: 10.14573/altex.1506201 PubMed.

(32) Van Norman, G. A. Drugs, Devices, and the FDA: Part 1: An Overview of Approval Processes for Drugs. *JACC Basic Transl Sci* **2016**, *1* (3), 170-179. DOI: 10.1016/j.jacbts.2016.03.002 PubMed.

(33) Lee, K. H.; Baik, S. Y.; Lee, S. Y.; Park, C. H.; Park, P. J.; Kim, J. H. Genome Sequence Variability Predicts Drug Precautions and Withdrawals from the Market. *PLoS One* **2016**, *11* (9), e0162135. DOI: 10.1371/journal.pone.0162135 PubMed.

(34) McNaughton, R.; Huet, G.; Shakir, S. An investigation into drug products withdrawn from the EU market between 2002 and 2011 for safety reasons and the evidence used to support the decision-making. *BMJ Open* **2014**, *4* (1), e004221. DOI: 10.1136/bmjopen-2013-004221 PubMed.

(35) Park, B. K.; Boobis, A.; Clarke, S.; Goldring, C. E. P.; Jones, D.; Kenna, J. G.; Lambert, C.; Laverty, H. G.; Naisbitt, D. J.; Nelson, S.; et al. Managing the challenge of chemically reactive metabolites in drug development. *Nature Reviews Drug Discovery* **2011**, *10* (4), 292-306. DOI: 10.1038/nrd3408.

(36) O'Brien, P. J.; Siraki, A. G.; Shangari, N. Aldehyde Sources, Metabolism, Molecular Toxicity Mechanisms, and Possible Effects on Human Health. *Critical Reviews in Toxicology* **2005**, *35* (7), 609-662. DOI: 10.1080/10408440591002183.

(37) Mak, I. W.; Evaniew, N.; Ghert, M. Lost in Translation: Animal Models and Clinical Trials in Cancer Treatment. *American Journal of Translational Research* **2014**, *6* (2), 114.

(38) Smietana, K.; Siatkowski, M.; Møller, M. Trends in clinical success rates. *Nature Reviews Drug Discovery* **2016**, *15* (6), 379-380. DOI: 10.1038/nrd.2016.85.

(39) Evens, R. P. Pharma Success in Product Development—Does Biotechnology Change the Paradigm in Product Development and Attrition. *AAPS J* **2016**, *18*, 281-285. DOI: 10.1208/s12248-015-9833-6 PubMed.

(40) Festing, S.; Wilkinson, R. The ethics of animal research: talking point on the use of animals in scientific research. *EMBO reports* **2007**, *8* (6), 526–530.

(41) Varga, O. E.; Hansen, A. K.; Sandøe, P.; Olsson, I. A. S. Validating Animal Models for Preclinical Research: A Scientific and Ethical Discussion. *ATLA Alternatives to Laboratory Animals* **2010**, *38* (3), 245–248.

(42) Adler, S.; Basketter, D.; Creton, S.; Pelkonen, O.; Van Benthem, J.; Zuang, V.; Andersen, K. E.; Angers-Loustau, A.; Aptula, A.; Bal-Price, A.; et al. Alternative (Non-Animal) Methods for Cosmetics Testing: Current Status and Future Prospects-2010. *Archives of Toxicology* **2011**, *85* (5), 367–485.

(43) Burden, N.; Mahony, C.; Müller, B. P.; Terry, C.; Westmoreland, C.; Kimber, I. Aligning the 3Rs with new paradigms in the safety assessment of chemicals. *Toxicology* **2015**, *330*, 62-66. DOI: 10.1016/j.tox.2015.01.014.

(44) Sullivan, K. M.; Manuppello, J. R.; Willett, C. E. Building on a solid foundation: SAR and QSAR as a fundamental strategy to reduce animal testing. *SAR and QSAR in Environmental Research* **2014**, *25* (5), 357-365. DOI: 10.1080/1062936x.2014.907203.

(45) Chapman, K. L.; Holzgrefe, H.; Black, L. E.; Brown, M.; Chellman, G.; Copeman, C.; Couch, J.; Creton, S.; Gehen, S.; Hoberman, A.; et al. Pharmaceutical toxicology: Designing studies to reduce animal use, while maximizing human translation. *Regulatory Toxicology and Pharmacology* **2013**, *66* (1), 88-103. DOI: 10.1016/j.yrtph.2013.03.001.

(46) Knudsen, T. B.; Keller, D. A.; Sander, M.; Carney, E. W.; Doerrer, N. G.; Eaton, D. L.; Fitzpatrick, S. C.; Hastings, K. L.; Mendrick, D. L.; Tice, R. R.; et al. FutureTox II: in vitro data and in silico models for predictive toxicology. *Toxicol Sci* **2015**, *143* (2), 256-267. DOI: 10.1093/toxsci/kfu234 PubMed.

(47) Li, X.; Chen, L.; Cheng, F.; Wu, Z.; Bian, H.; Xu, C.; Li, W.; Liu, G.; Shen, X.; Tang, Y. In Silico Prediction of Chemical Acute Oral Toxicity Using Multi-Classification Methods. *Journal of Chemical Information and Modeling* **2014**, *54* (4), 1061-1069. DOI: 10.1021/ci5000467.

(48) Goh, G. B.; Hodas, N. O.; Vishnu, A. Deep learning for computational chemistry. *Journal of Computational Chemistry* **2017**, *38* (16), 1291-1307. DOI: 10.1002/jcc.24764.

(49) Raies, A. B.; Bajic, V. B. In silico toxicology: computational methods for the prediction of chemical toxicity. *Wiley Interdiscip Rev Comput Mol Sci* **2016**, *6* (2), 147-172. DOI: 10.1002/wcms.1240 PubMed.

(50) Raunio, H. In silico toxicology - non-testing methods. *Front Pharmacol* **2011**, *2*, 33. DOI: 10.3389/fphar.2011.00033 PubMed.

(51) Greene, N.; Judson, P. N.; Langowski, J. J.; Marchant, C. A. Knowledge-Based Expert Systems for Toxicity and Metabolism Prediction: DEREK, StAR and METEOR. *SAR and QSAR in Environmental Research* **1999**, *10* (2-3), 299-314. DOI: 10.1080/10629369908039182.

(52) Marchant, C. A.; Briggs, K. A.; Long, A. In Silico Tools for Sharing Data and Knowledge on Toxicity and Metabolism: Derek for Windows, Meteor, and Vitic. *Toxicology Mechanisms and Methods* **2008**, *18* (2-3), 177-187. DOI: 10.1080/15376510701857320.

(53) Mombelli, E.; Devillers, J. Evaluation of the OECD (Q)SAR Application Toolbox and Toxtree for predicting and profiling the carcinogenic potential of chemicals. *SAR and QSAR in Environmental Research* **2010**, *21* (7-8), 731-752. DOI: 10.1080/1062936x.2010.528598.

(54) Dimitrov, S. D.; Diderich, R.; Sobanski, T.; Pavlov, T. S.; Chankov, G. V.; Chapkanov, A. S.; Karakolev, Y. H.; Temelkov, S. G.; Vasilev, R. A.; Gerova, K. D.; et al. QSAR Toolbox

– workflow and major functionalities. *SAR and QSAR in Environmental Research* **2016**, *27* (3), 203-219. DOI: 10.1080/1062936x.2015.1136680.

(55) Yang, H.; Sun, L.; Li, W.; Liu, G.; Tang, Y. Corrigendum: In Silico Prediction of Chemical Toxicity for Drug Design Using Machine Learning Methods and Structural Alerts. *Front Chem* **2018**, *6*, 129. DOI: 10.3389/fchem.2018.00129 PubMed.

(56) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; et al. Applications of Machine Learning in Drug Discovery and Development. *Nature Reviews Drug Discovery* **2019**, *18* (6), 463–477.

(57) Baskin, I. I. Machine Learning Methods in Computational Toxicology. In *Computational Toxicology.*, Nicolotti, O. Ed.; Humana Press, New York, NY, 2018; pp 119-139.

(58) Villeneuve, D. L.; Crump, D.; Garcia-Reyero, N.; Hecker, M.; Hutchinson, T. H.; LaLone, C. A.; Landesmann, B.; Lettieri, T.; Munn, S.; Nepelska, M.; et al. Adverse outcome pathway (AOP) development I: strategies and principles. *Toxicol Sci* **2014**, *142* (2), 312-320. DOI: 10.1093/toxsci/kfu199 PubMed.

(59) Knapen, D.; Vergauwen, L.; Villeneuve, D. L.; Ankley, G. T. The potential of AOP networks for reproductive and developmental toxicity assay development. *Reproductive Toxicology* **2015**, *56*, 52-55. DOI: 10.1016/j.reprotox.2015.04.003.

(60) Richard, A. M.; Judson, R. S.; Houck, K. A.; Grulke, C. M.; Volarath, P.; Thillainadarajah, I.; Yang, C.; Rathman, J.; Martin, M. T.; Wambaugh, J. F.; et al. ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chemical Research in Toxicology* **2016**, *29* (8), 1225-1251. DOI: 10.1021/acs.chemrestox.6b00135.

(61) Dix, D. J.; Houck, K. A.; Martin, M. T.; Richard, A. M.; Setzer, R. W.; Kavlock, R. J. The ToxCast Program for Prioritizing Toxicity Testing of Environmental Chemicals. *Toxicol Sci* **2007**, *95* (1), 5-12. DOI: 10.1093/toxsci/kfl103.

(62) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; et al. The ChEMBL database in 2017. *Nucleic Acids Res* **2017**, *45* (D1), D945-D954. DOI: 10.1093/nar/gkw1074 PubMed.

(63) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* **2012**, *40* (D1), D1100-D1107. DOI: 10.1093/nar/gkr777 PubMed.

(64) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; et al. The ChEMBL bioactivity database: an update.

Nucleic Acids Res **2014**, *42* (Database issue), D1083-D1090. DOI: 10.1093/nar/gkt1031 PubMed.

(65) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2019**, *47* (D1), D1102-D1109.

(66) Bowes, J.; Brown, A. J.; Hamon, J.; Jarolimek, W.; Sridhar, A.; Waldron, G.; Whitebread, S. Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nature Reviews Drug Discovery* **2012**, *11* (12), 909-922. DOI: 10.1038/nrd3845.

(67) Schulz, M.; Schmoldt, A.; Schulz, M. Therapeutic and Toxic Blood Concentrations of More than 800 Drugs and Other Xenobiotics. *Die Pharmazie-An International Journal of Pharmaceutical Sciences* **2003**, *58* (7), 447-474.

(68) Sagar, S.; Kaur, M.; Radovanovic, A.; Bajic, V. B. Dragon exploration system on marine sponge compounds interactions. *J Cheminform* **2013**, *5* (1), 1-7. DOI: 10.1186/1758-2946-5-11 PubMed.

(69) Cases, M.; Pastor, M.; Sanz, F. The eTOX Library of Public Resources for in Silico Toxicity Prediction. *Molecular Informatics* **2013**, *32* (1), 24-35. DOI: 10.1002/minf.201200099.

(70) Wishart, D. S.; Knox, C.; Guo, A. C.; Eisner, R.; Young, N.; Gautam, B.; Hau, D. D.; Psychogios, N.; Dong, E.; Bouatra, S.; et al. HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res* **2009**, *37* (suppl_1), D603-D610. DOI: 10.1093/nar/gkn810 PubMed.

(71) Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; et al. HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic Acids Res* **2012**, *41* (D1), D801-D807. DOI: 10.1093/nar/gks1065 PubMed.

(72) Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; et al. HMDB: the Human Metabolome Database. *Nucleic Acids Res* **2007**, *35* (suppl_1), D521-D526. DOI: 10.1093/nar/gkl923 PubMed.

(73) Han, H.; Wang, W.-Y.; Mao, B.-H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In *Lecture Notes in Computer Science*, Huang, D., Zhang, XP., Huang, GB. Ed.; Vol. 3644; Springer Berlin Heidelberg, 2005; pp 878-887.

(74) Chen, J. J.; Tsai, C. A.; Young, J. F.; Kodell, R. L. Classification ensembles for unbalanced class sizes in predictive toxicology. *SAR and QSAR in Environmental Research* **2005**, *16* (6), 517-529. DOI: 10.1080/10659360500468468.

(75) Fernandez, A.; Garcia, S.; Herrera, F.; Chawla, N. V. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research* **2018**, *61*, 863-905. DOI: 10.1613/jair.1.11192.

(76) Buda, M.; Maki, A.; Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks* **2018**, *106*, 249-259. DOI: 10.1016/j.neunet.2018.07.011.

(77) Lin, W.-C.; Tsai, C.-F.; Hu, Y.-H.; Jhang, J.-S. Clustering-based undersampling in class-imbalanced data. *Information Sciences* **2017**, *409*, 17-26. DOI: 10.1016/j.ins.2017.05.008.

(78) Boughorbel, S.; Jarray, F.; El-Anbari, M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS One* **2017**, *12* (6), e0177678. DOI: 10.1371/journal.pone.0177678 PubMed.

(79) Yan, Y.; Chen, M.; Shyu, M.-L.; Chen, S.-C. Deep Learning for Imbalanced Multimedia Data Classification. In 2015 IEEE International Symposium on Multimedia (ISM), 2015.

(80) Akbani, R.; Kwek, S.; Japkowicz, N. Applying Support Vector Machines to Imbalanced Datasets. In *Machine Learning: ECML 2004: 15th European Conference on Machine Learning*, Pisa, Italy, 2004; Springer Berlin Heidelberg: Vol. Proceedings 15, pp 39-50. DOI: 10.1007/978-3-540-30115-8_7.

(81) Batista, G. E.; Prati, R. C.; Monard, M. C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter* **2004**, *6* (1), 20-29. DOI: 10.1145/1007730.1007735.

(82) Lemaître, G.; Nogueira, F.; Aridas, C. K. Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *The Journal of Machine Learning Research* **2017**, *18* (1), 559-563.

(83) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J Chem Inf Model* **2010**, *50* (7), 1189-1204. DOI: 10.1021/ci100176x From NLM.

(84) Gimadiev, T. R.; Lin, A.; Afonina, V. A.; Batyrshin, D.; Nugmanov, R. I.; Akhmetshin, T.; Sidorov, P.; Duybankova, N.; Verhoeven, J.; Wegner, J.; et al. Reaction data curation I: chemical structures and transformations standardization. *Molecular Informatics* **2021**, *40* (12), 2100119.

(85) Erickson, B. J.; Korfiatis, P.; Akkus, Z.; Kline, T. L. Machine Learning for Medical Imaging. *Radiographics* **2017**, *37* (2), 505-515. DOI: 10.1148/rg.2017160130 PubMed.

(86) Beam, A. L.; Kohane, I. S. Big Data and Machine Learning in Health Care. *JAMA* **2018**, *319* (13), 1317-1318. DOI: 10.1001/jama.2017.18391.

(87) Gawehn, E.; Hiss, J. A.; Schneider, G. Deep Learning in Drug Discovery. *Molecular Informatics* **2016**, *35* (1), 3-14. DOI: 10.1002/minf.201501008.

(88) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discovery Today* **2018**, *23* (6), 1241–1250.

(89) Nichols, J. A.; Herbert Chan, H. W.; Baker, M. A. Machine learning: applications of artificial intelligence to imaging and diagnosis. *Biophys Rev* **2019**, *11*, 111-118. DOI: 10.1007/s12551-018-0449-9 PubMed.

(90) Wang, X.; Wang, X.; Wilkes, D. M. Supervised Learning for Data Classification Based Object Recognition. In *Machine Learning-based Natural Scene Recognition for Mobile Robot Localization in An Unknown Environment*, Springer Singapore, 2020; pp 179-194.

(91) Jordan, M. I.; Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349* (6245), 255-260. DOI: 10.1126/science.aaa8415.

(92) Schrider, D. R.; Kern, A. D. Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends Genet* **2018**, *34* (4), 301-312. DOI: 10.1016/j.tig.2017.12.005 PubMed.

(93) Libbrecht, M. W.; Noble, W. S. Machine learning applications in genetics and genomics. *Nat Rev Genet* **2015**, *16* (6), 321-332. DOI: 10.1038/nrg3920 PubMed.

(94) Längkvist, M.; Karlsson, L.; Loutfi, A. A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters* **2014**, *42*, 11-24. DOI: 10.1016/j.patrec.2014.01.008.

(95) Le, Q. V. Building high-level features using large scale unsupervised learning. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.

(96) Raina, R.; Madhavan, A.; Ng, A. Y. Large-scale deep unsupervised learning using graphics processors. In *Proceedings of the 26th annual international conference on machine learning*, 2009; ACM Press: pp 873-880. DOI: 10.1145/1553374.1553486.

(97) van Engelen, J. E.; Hoos, H. H. A survey on semi-supervised learning. *Machine Learning* **2020**, *109* (2), 373-440. DOI: 10.1007/s10994-019-05855-6.

(98) Guo, B.; Tao, H.; Hou, C.; Yi, D. Semi-supervised multi-label feature learning via label enlarged discriminant analysis. *Knowledge and Information Systems* **2020**, *62*, 2383-2417. DOI: 10.1007/s10115-019-01409-3.

(99) Kostopoulos, G.; Kotsiantis, S.; Fazakis, N.; Koutsonikos, G.; Pierrakeas, C. A Semi-Supervised Regression Algorithm for Grade Prediction of Students in Distance Learning Courses. *International Journal on Artificial Intelligence Tools* **2019**, *28* (04), 1940001. DOI: 10.1142/s0218213019400013.

(100) Zhou, L.; Liu, Z.; Tan, H.; Xie, X. Semisupervised learning with adversarial training among joint distributions. *Journal of Electronic Imaging* **2019**, *28* (5), 053030. DOI: 10.1117/1.jei.28.5.053030.

(101) Wang, J.; Zuo, R.; Xiong, Y. Mapping Mineral Prospectivity via Semi-supervised Random Forest. *Natural Resources Research* **2020**, *29*, 189-202. DOI: 10.1007/s11053-019-09510-8.

(102) Huang, G.; Song, S.; Gupta, J. N. D.; Wu, C. Semi-Supervised and Unsupervised Extreme Learning Machines. *IEEE Trans. Cybern.* **2014**, *44* (12), 2405–2417.

(103) Tanha, J.; van Someren, M.; Afsarmanesh, H. Semi-supervised self-training for decision tree classifiers. *International Journal of Machine Learning and Cybernetics* **2017**, *8*, 355-370. DOI: 10.1007/s13042-015-0328-7.

(104) Zhou, Z.-H. A brief introduction to weakly supervised learning. *National Science Review* **2018**, *5* (1), 44-53. DOI: 10.1093/nsr/nwx106.

(105) Idakwo, G.; Thangapandian, S.; Luttrell, J. t.; Zhou, Z.; Zhang, C.; Gong, P. 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* **2019**, *10*, 1044. DOI: 10.3389/fphys.2019.01044 PubMed.

(106) Sun, L.; Yang, H.; Cai, Y.; Li, W.; Liu, G.; Tang, Y. In Silico Prediction of Endocrine Disrupting Chemicals Using Single-Label and Multilabel Models. *Journal of Chemical Information and Modeling* **2019**, *59* (3), 973-982. DOI: 10.1021/acs.jcim.8b00551.

(107) 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.

(108) Xu, Y.; Pei, J.; Lai, L. Deep Learning Based Regression and Multiclass Models for Acute Oral Toxicity Prediction with Automatic Chemical Feature Extraction. *J. Chem. Inf. Model.* **2017**, *57* (11), 2672–2685.

(109) Pereira, J. C.; Caffarena, E. R.; dos Santos, C. N. Boosting Docking-Based Virtual Screening with Deep Learning. *Journal of Chemical Information and Modeling* **2016**, *56* (12), 2495-2506. DOI: 10.1021/acs.jcim.6b00355.

(110) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50* (5), 742-754. DOI: 10.1021/ci100050t.

(111) Cai, Y.; Yang, H.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. Computational Prediction of Site of Metabolism for UGT-Catalyzed Reactions. *Journal of Chemical Information and Modeling* **2018**, *59* (3), 1085-1095. DOI: 10.1021/acs.jcim.8b00851.

(112) Wang, Y.; Guo, Y.; Kuang, Q.; Pu, X.; Ji, Y.; Zhang, Z.; Li, M. A comparative study of family-specific protein–ligand complex affinity prediction based on random forest approach. *Journal of Computer-Aided Molecular Design* **2015**, *29*, 349-360. DOI: 10.1007/s10822-014-9827-y.

(113) Bartlett, M. S.; Littlewort, G.; Frank, M.; Lainscsek, C.; Fasel, I.; Movellan, J. Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005.

(114) Breiman, L. Random Forests. *Machine Learning* **2001**, *45* (1), 5-32. DOI: 10.1023/a:1010933404324.

(115) Kumari, P.; Nath, A.; Chaube, R. Identification of human drug targets using machine-learning algorithms. *Computers in Biology and Medicine* **2015**, *56*, 175-181. DOI: 10.1016/j.compbiomed.2014.11.008.

(116) Fooshee, D.; Mood, A.; Gutman, E.; Tavakoli, M.; Urban, G.; Liu, F.; Huynh, N.; Van Vranken, D.; Baldi, P. Deep learning for chemical reaction prediction. *Molecular Systems Design & Engineering* **2018**, *3* (3), 442-452. DOI: 10.1039/c7me00107j.

(117) Patil, P. M.; Suralkar, S. R.; Abhyankar, H. K. Fingerprint verification based on fixed length square finger code. In 17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'05), 2005.

(118) Jha, D.; Ward, L.; Paul, A.; Liao, W.-K.; Choudhary, A.; Wolverton, C.; Agrawal, A. ElemNet: Deep Learning the Chemistry of Materials From Only Elemental Composition. *Sci Rep* **2018**, *8* (1), 1-13. DOI: 10.1038/s41598-018-35934-y PubMed.

(119) Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Computers & Electrical Engineering* **2014**, *40* (1), 16-28. DOI: 10.1016/j.compeleceng.2013.11.024.

(120) Blum, A. L.; Langley, P. Selection of relevant features and examples in machine learning. *Artificial Intelligence* **1997**, *97* (1-2), 245-271. DOI: 10.1016/s0004-3702(97)00063-5.

(121) Korkmaz, S.; Zararsiz, G.; Goksuluk, D. Drug/nondrug classification using Support Vector Machines with various feature selection strategies. *Computer Methods and Programs in Biomedicine* **2014**, *117* (2), 51-60. DOI: 10.1016/j.cmpb.2014.08.009.

(122) Zhang, P.; Wang, F.; Hu, J.; Sorrentino, R. Label Propagation Prediction of Drug-Drug Interactions Based on Clinical Side Effects. *Sci Rep* **2015**, *5* (1), 12339. DOI: 10.1038/srep12339 PubMed.

(123) Rao, R. B.; Fung, G.; Rosales, R. On the Dangers of Cross-Validation. An Experimental Evaluation. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, 2008/04/24, 2008; Society for Industrial and Applied Mathematics: pp 588-596. DOI: 10.1137/1.9781611972788.54.

(124) Tan, N.-X.; Li, P.; Rao, H.-B.; Li, Z.-R.; Li, X.-Y. Prediction of the acute toxicity of chemical compounds to the fathead minnow by machine learning approaches. *Chemometrics and Intelligent Laboratory Systems* **2010**, *100* (1), 66-73. DOI: 10.1016/j.chemolab.2009.11.002.

(125) Wen, M.; Zhang, Z.; Niu, S.; Sha, H.; Yang, R.; Yun, Y.; Lu, H. Deep-Learning-Based Drug–Target Interaction Prediction. *Journal of Proteome Research* **2017**, *16* (4), 1401-1409. DOI: 10.1021/acs.jproteome.6b00618.

(126) Gayvert, K. M.; Madhukar, N. S.; Elemento, O. A Data-Driven Approach to Predicting Successes and Failures of Clinical Trials. *Cell Chem Biol* **2016**, *23* (10), 1294-1301. DOI: 10.1016/j.chembiol.2016.07.023 PubMed.

(127) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving beyond Fingerprints. *J. Comput. Aided. Mol. Des.* **2016**, *30*, 595–608.

(128) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular Fingerprint Similarity Search in Virtual Screening. *Methods* **2015**, *71*, 58–63.

(129) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273–1280.

(130) Schaffer, C. Selecting a classification method by cross-validation. *Machine Learning* **1993**, *13*, 135-143. DOI: 10.1007/bf00993106.

(131) Kerns, S. L.; Kundu, S.; Oh, J. H.; Singhal, S. K.; Janelsins, M.; Travis, L. B.; Deasy, J. O.; Janssens, A. C. J. E.; Ostrer, H.; Parliament, M.; et al. The Prediction of Radiotherapy Toxicity Using Single Nucleotide Polymorphism-Based Models: A Step Toward Prevention. *Semin Radiat Oncol* **2015**, *25* (4), 281-291. DOI: 10.1016/j.semradonc.2015.05.006 PubMed.

(132) Linden, A.; Yarnold, P. R.; Nallamothu, B. K. Using machine learning to model dose-response relationships. *Journal of Evaluation in Clinical Practice* **2016**, *22* (6), 860-867. DOI: 10.1111/jep.12573.

(133) Zhong, E.; Fan, W.; Yang, Q.; Verscheure, O.; Ren, J. Cross Validation Framework to Choose amongst Models and Datasets for Transfer Learning. In *Lecture Notes in Computer*

Science, Balcázar, J. L., Bonchi, F., Gionis, A., Sebag, M. Ed.; Vol. 6323; Springer, Berlin, Heidelberg, 2010; pp 547-562.

(134) Cawley, G. C.; Talbot, N. L. Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. *Pattern Recognition* **2003**, *36* (11), 2585-2592. DOI: 10.1016/s0031-3203(03)00136-5.

(135) Cawley, G. C.; Talbot, N. L. Fast exact leave-one-out cross-validation of sparse least-squares support vector machines. *Neural Networks* **2004**, *17* (10), 1467-1475. DOI: 10.1016/j.neunet.2004.07.002.

(136) Chen, T.; Cao, Y.; Zhang, Y.; Liu, J.; Bao, Y.; Wang, C.; Jia, W.; Zhao, A. Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection. *Evid Based Complement Alternat Med* **2013**, *2013*, 298183. DOI: 10.1155/2013/298183 PubMed.

(137) He, H.; Bai, Y.; Garcia, E. A.; Li, S. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), 2008.

(138) Graser, J.; Kauwe, S. K.; Sparks, T. D. Machine Learning and Energy Minimization Approaches for Crystal Structure Predictions: A Review and New Horizons. *Chemistry of Materials* **2018**, *30* (11), 3601-3612. DOI: 10.1021/acs.chemmater.7b05304.

(139) An, S.; Liu, W.; Venkatesh, S. Fast Cross-Validation Algorithms for Least Squares Support Vector Machine and Kernel Ridge Regression. *Pattern Recognit.* **2007**, *40* (8), 2154–2162.

(140) Rácz, A.; Bajusz, D.; Héberger, K. SAR and QSAR in Environmental Research Modelling Methods and Cross-Validation Variants in QSAR: A Multi-Level Analysis. *SAR QSAR Environ. Res.* **2018**, *29* (9), 661–674.

(141) Xu, Q. S.; Liang, Y. Z. Monte Carlo cross validation. *Chemom. Intell. Lab. Syst.* **2001**, *56* (1), 1–11.

(142) He, S.; Ye, T.; Wang, R.; Zhang, C.; Zhang, X.; Sun, G.; Sun, X. An In Silico Model for Predicting Drug-Induced Hepatotoxicity. *Int J Mol Sci* **2019**, *20* (8), 1897. DOI: 10.3390/ijms20081897 PubMed.

(143) Liu, J.; Mansouri, K.; Judson, R. S.; Martin, M. T.; Hong, H.; Chen, M.; Xu, X.; Thomas, R. S.; Shah, I. Predicting Hepatotoxicity Using ToxCast in Vitro Bioactivity and Chemical Structure. *Chemical Research in Toxicology* **2015**, *28* (4), 738-751. DOI: 10.1021/tx500501h.

(144) Liu, J.; Patlewicz, G.; Williams, A. J.; Thomas, R. S.; Shah, I. Predicting Organ Toxicity Using in Vitro Bioactivity Data and Chemical Structure. *Chem. Res. Toxicol.* **2017**, *30* (11), 2046–2059.

(145) Zhang, H.; Cao, Z.-X.; Li, M.; Li, Y.-Z.; Peng, C. Novel naïve Bayes classification models for predicting the carcinogenicity of chemicals. *Food and Chemical Toxicology* **2016**, *97*, 141-149. DOI: 10.1016/j.fct.2016.09.005.

(146) Zhang, H.; Ren, J.-X.; Kang, Y.-L.; Bo, P.; Liang, J.-Y.; Ding, L.; Kong, W.-B.; Zhang, J. Development of novel in silico model for developmental toxicity assessment by using naïve Bayes classifier method. *Reproductive Toxicology* **2017**, *71*, 8-15. DOI: 10.1016/j.reprotox.2017.04.005.

(147) Zhang, H.; Ding, L.; Zou, Y.; Hu, S.-Q.; Huang, H.-G.; Kong, W.-B.; Zhang, J. Predicting drug-induced liver injury in human with Naïve Bayes classifier approach. *Journal of Computer-Aided Molecular Design* **2016**, *30*, 889-898. DOI: 10.1007/s10822-016-9972-6.

(148) Schrey, A. K.; Nickel-Seeber, J.; Drwal, M. N.; Zwicker, P.; Schultze, N.; Haertel, B.; Preissner, R. Computational Prediction of Immune Cell Cytotoxicity. *Food Chem. Toxicol.* **2017**, *107*, 150–166.

(149) Zhang, H.; Kang, Y.-L.; Zhu, Y.-Y.; Zhao, K.-X.; Liang, J.-Y.; Ding, L.; Zhang, T.-G.; Zhang, J. Novel naïve Bayes classification models for predicting the chemical Ames mutagenicity. *Toxicology in Vitro* **2017**, *41*, 56-63. DOI: 10.1016/j.tiv.2017.02.016.

(150) Seal, A.; Passi, A.; Jaleel, U. A.; Consortium, O. S. D. D.; Wild, D. J. In-silico predictive mutagenicity model generation using supervised learning approaches. *J Cheminform* **2012**, *4*, 1-11. DOI: 10.1186/1758-2946-4-10 PubMed.

(151) Zhang, H.; Yu, P.; Ren, J.-X.; Li, X.-B.; Wang, H.-L.; Ding, L.; Kong, W.-B. Development of novel prediction model for drug-induced mitochondrial toxicity by using naïve Bayes classifier method. *Food and Chemical Toxicology* **2017**, *110*, 122-129. DOI: 10.1016/j.fct.2017.10.021.

(152) Zhang, H.; Yu, P.; Zhang, T.-G.; Kang, Y.-L.; Zhao, X.; Li, Y.-Y.; He, J.-H.; Zhang, J. In silico prediction of drug-induced myelotoxicity by using Naïve Bayes method. *Molecular Diversity* **2015**, *19*, 945-953. DOI: 10.1007/s11030-015-9613-3.

(153) Su, R.; Li, Y.; Zink, D.; Loo, L. H. Supervised prediction of drug-induced nephrotoxicity based on interleukin-6 and -8 expression levels. *BMC Bioinformatics* **2014**, *15* (16), 1-9. DOI: 10.1186/1471-2105-15-S16-S16 PubMed.

(154) Zhang, H.; Ma, J.-X.; Liu, C.-T.; Ren, J.-X.; Ding, L. Development and evaluation of in silico prediction model for drug-induced respiratory toxicity by using naïve Bayes classifier method. *Food and Chemical Toxicology* **2018**, *121*, 593-603. DOI: 10.1016/j.fct.2018.09.051.

(155) Zhang, H.; Ren, J.-X.; Ma, J.-X.; Ding, L. Development of an in silico prediction model for chemical-induced urinary tract toxicity by using naïve Bayes classifier. *Molecular Diversity* **2019**, *23*, 381-392. DOI: 10.1007/s11030-018-9882-8.

(156) Pella, A.; Cambria, R.; Riboldi, M.; Jereczek-Fossa, B. A.; Fodor, C.; Zerini, D.; Torshabi, A. E.; Cattani, F.; Garibaldi, C.; Pedroli, G.; et al. Use of Machine Learning Methods for Prediction of Acute Toxicity in Organs at Risk Following Prostate Radiotherapy. *Med. Phys.* **2011**, *38* (6Part1), 2859–2867.

(157) Zhang, L.; Ai, H.; Chen, W.; Yin, Z.; Hu, H.; Zhu, J.; Zhao, J.; Zhao, Q.; Liu, H. CarcinoPred-EL: Novel models for predicting the carcinogenicity of chemicals using molecular fingerprints and ensemble learning methods. *Sci Rep* **2017**, *7* (1), 2118. DOI: 10.1038/s41598-017-02365-0 PubMed.

(158) Shen, M.-y.; Su, B.-H.; Esposito, E. X.; Hopfinger, A. J.; Tseng, Y. J. A Comprehensive Support Vector Machine Binary hERG Classification Model Based on Extensive but Biased End Point hERG Data Sets. *Chemical Research in Toxicology* **2011**, *24* (6), 934-949. DOI: 10.1021/tx200099j.

(159) Li, X.; Chen, Y.; Song, X.; Zhang, Y.; Li, H.; Zhao, Y. The development and application of in silico models for drug induced liver injury. *RSC Advances* **2018**, *8* (15), 8101-8111. DOI: 10.1039/c7ra12957b.

(160) Sharma, A.; Kumar, R.; Varadwaj, P. K.; Ahmad, A.; Ashraf, G. M. A comparative study of support vector machine, artificial neural network and Bayesian classifier for mutagenicity prediction. *Interdisciplinary Sciences: Computational Life Sciences* **2011**, *3*, 232-239. DOI: 10.1007/s12539-011-0102-9.

(161) Zhou, S.; Li, G.-B.; Huang, L.-Y.; Xie, H.-Z.; Zhao, Y.-L.; Chen, Y.-Z.; Li, L.-L.; Yang, S.-Y. A prediction model of drug-induced ototoxicity developed by an optimal support vector machine (SVM) method. *Computers in Biology and Medicine* **2014**, *51*, 122-127. DOI: 10.1016/j.compbiomed.2014.05.005.

(162) Lee, H.-M.; Yu, M.-S.; Kazmi, S. R.; Oh, S. Y.; Rhee, K.-H.; Bae, M.-A.; Lee, B. H.; Shin, D.-S.; Oh, K.-S.; Ceong, H.; et al. Computational determination of hERG-related cardiotoxicity of drug candidates. *BMC Bioinformatics* **2019**, *20*, 67-73. DOI: 10.1186/s12859-019-2814-5 PubMed.

(163) Siramshetty, V. B.; Chen, Q.; Devarakonda, P.; Preissner, R. The Catch-22 of Predicting hERG Blockade Using Publicly Accessible Bioactivity Data. *Journal of Chemical Information and Modeling* **2018**, *58* (6), 1224-1233. DOI: 10.1021/acs.jcim.8b00150.

(164) Zhang, Y.; Zhao, J.; Wang, Y.; Fan, Y.; Zhu, L.; Yang, Y.; Chen, X.; Lu, T.; Chen, Y.; Liu, H. Prediction of HERG K+ Channel Blockage Using Deep Neural Networks. *Chem. Biol. Drug Des.* **2019**, *94* (5), 1973–1985.

(165) Kim, E.; Nam, H. Prediction models for drug-induced hepatotoxicity by using weighted molecular fingerprints. *BMC Bioinformatics* **2017**, *18*, 25-34. DOI: 10.1186/s12859-017-1638-4 PubMed.

(166) Ai, H.; Chen, W.; Zhang, L.; Huang, L.; Yin, Z.; Hu, H.; Zhao, Q.; Zhao, J.; Liu, H. Predicting Drug-Induced Liver Injury Using Ensemble Learning Methods and Molecular Fingerprints. *Toxicol Sci* **2018**, *165* (1), 100-107. DOI: 10.1093/toxsci/kfy121.

(167) Kandasamy, K.; Chuah, J. K. C.; Su, R.; Huang, P.; Eng, K. G.; Xiong, S.; Li, Y.; Chia, C. S.; Loo, L. H.; Zink, D. Prediction of Drug-Induced Nephrotoxicity and Injury Mechanisms with Human Induced Pluripotent Stem Cell-Derived Cells and Machine Learning Methods. *Sci. Rep.* **2015**, *5* (1), 12337.

(168) Cai, C.; Guo, P.; Zhou, Y.; Zhou, J.; Wang, Q.; Zhang, F.; Fang, J.; Cheng, F. Deep Learning-Based Prediction of Drug-Induced Cardiotoxicity. *Journal of chemical information and modeling* **2019**, *59* (3), 1073-1084. DOI: 10.1021/acs.jcim.8b00769 PubMed.

(169) Xu, Y.; Dai, Z.; Chen, F.; Gao, S.; Pei, J.; Lai, L. Deep Learning for Drug-Induced Liver Injury. *Journal of Chemical Information and Modeling* **2015**, *55* (10), 2085-2093. DOI: 10.1021/acs.jcim.5b00238.

(170) Xu, C.; Cheng, F.; Chen, L.; Du, Z.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. In silico Prediction of Chemical Ames Mutagenicity. *Journal of Chemical Information and Modeling* **2012**, *52* (11), 2840-2847. DOI: 10.1021/ci300400a.

(171) Fernandez, M.; Ban, F.; Woo, G.; Hsing, M.; Yamazaki, T.; LeBlanc, E.; Rennie, P. S.; Welch, W. J.; Cherkasov, A. Toxic Colors: The Use of Deep Learning for Predicting Toxicity of Compounds Merely from Their Graphic Images. *Journal of Chemical Information and Modeling* **2018**, *58* (8), 1533-1543. DOI: 10.1021/acs.jcim.8b00338.

(172) Yuan, Q.; Wei, Z.; Guan, X.; Jiang, M.; Wang, S.; Zhang, S.; Li, Z. Toxicity Prediction Method Based on Multi-Channel Convolutional Neural Network. *Molecules (Basel, Switzerland)* **2019**, *24* (18), 3383. DOI: 10.3390/molecules24183383 PubMed.

(173) Sharma, A. K.; Srivastava, G. N.; Roy, A.; Sharma, V. K. ToxiM: A Toxicity Prediction Tool for Small Molecules Developed Using Machine Learning and Chemoinformatics Approaches. *Front. Pharmacol.* **2017**, *8*, 880.

(174) Su, R.; Wu, H.; Liu, X.; Wei, L. Predicting drug-induced hepatotoxicity based on biological feature maps and diverse classification strategies. *Briefings in Bioinformatics* **2021**, *22* (1), 428-437. DOI: 10.1093/bib/bbz165.

(175) Yajima, D.; Ohkawa, T.; Muroi, K.; Imaishi, H. Predicting Toxicity of Food-Related Compounds Using Fuzzy Decision Trees. *Int. J. Biosci. Biochem. Bioinforma.* **2014**, *4* (1), 33.

(176) Hammann, F.; Schöning, V.; Drewe, J. Prediction of Clinically Relevant Drug-Induced Liver Injury from Structure Using Machine Learning. *J. Appl. Toxicol.* **2019**, *39* (3), 412–419.

(177) Li, F.; Fan, D.; Wang, H.; Yang, H.; Li, W.; Tang, Y.; Liu, G. In Silico Prediction of Pesticide Aquatic Toxicity with Chemical Category Approaches. *Toxicology research* **2017**, *6* (6), 831–842.

(178) Li, X.; Du, Z.; Wang, J.; Wu, Z.; Li, W.; Liu, G.; Shen, X.; Tang, Y. In Silico Estimation of Chemical Carcinogenicity with Binary and Ternary Classification Methods. *Molecular Informatics* **2015**, *34* (4), 228-235. DOI: 10.1002/minf.201400127.

(179) Zhang, C.; Zhou, Y.; Gu, S.; Wu, Z.; Wu, W.; Liu, C.; Wang, K.; Liu, G.; Li, W.; Lee, P. W.; et al. In silico prediction of hERG potassium channel blockage by chemical category approaches. *Toxicology research* **2016**, *5* (2), 570-582. DOI: 10.1039/c5tx00294j PubMed.

(180) Fan, D.; Yang, H.; Li, F.; Sun, L.; Di, P.; Li, W.; Tang, Y.; Liu, G. In silico prediction of chemical genotoxicity using machine learning methods and structural alerts. *Toxicology research* **2018**, *7* (2), 211-220. DOI: 10.1039/c7tx00259a PubMed.

(181) Zhang, C.; Cheng, F.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. In silico Prediction of Drug Induced Liver Toxicity Using Substructure Pattern Recognition Method. *Molecular Informatics* **2016**, *35* (3-4), 136-144. DOI: 10.1002/minf.201500055.

(182) Richard, A. M.; Huang, R.; Waidyanatha, S.; Shinn, P.; Collins, B. J.; Thillainadarajah, I.; Grulke, C. M.; Williams, A. J.; Lougee, R. R.; Judson, R. S.; et al. The Tox21 10K Compound Library: Collaborative Chemistry Advancing Toxicology. *Chem. Res. Toxicol.* **2020**, *34* (2), 189-216.

(183) Hsieh, J.-H.; Smith-Roe, S. L.; Huang, R.; Sedykh, A.; Shockley, K. R.; Auerbach, S. S.; Merrick, B. A.; Xia, M.; Tice, R. R.; Witt, K. L. Identifying Compounds with Genotoxicity Potential Using Tox21 High-Throughput Screening Assays. *Chemical research in toxicology* **2019**, *32* (7), 1384–1401.

(184) Huang, R.; Xia, M.; Sakamuru, S.; Zhao, J.; Shahane, S. A.; Attene-Ramos, M.; Zhao, T.; Austin, C. P.; Simeonov, A. Modelling the Tox21 10 K Chemical Profiles for in Vivo Toxicity Prediction and Mechanism Characterization. *Nat. Commun.* **2016**, *7* (1), 10425.

(185) Capuzzi, S. J.; Politi, R.; Isayev, O.; Farag, S.; Tropsha, A. QSAR Modeling of Tox21 Challenge Stress Response and Nuclear Receptor Signaling Toxicity Assays. *Front. Environ. Sci.* **2016**, *4*, 3.

(186) Judson, R.; Houck, K.; Martin, M.; Knudsen, T.; Thomas, R. S.; Sipes, N.; Shah, I.; Wambaugh, J.; Crofton, K. In Vitro and Modelling Approaches to Risk Assessment from the U.S. Environmental Protection Agency ToxCast Programme. *Basic Clin. Pharmacol. Toxicol.* **2014**, *115* (1), 69–76.

(187) Norinder, U.; Boyer, S. Conformal Prediction Classification of a Large Data Set of EnRvironmental Chemicals from ToxCast and Tox21 Estrogen Receptor Assays. *Chem. Res. Toxicol.* **2016**, *29* (6), 1003–1010.

(188) Sipes, N. S.; Martin, M. T.; Reif, D. M.; Kleinstreuer, N. C.; Judson, R. S.; Singh, A. V.; Chandler, K. J.; Dix, D. J.; Kavlock, R. J.; Knudsen, T. B. Predictive Models of Prenatal Developmental Toxicity from Toxcast High-Throughput Screening Data. *Toxicol. Sci.* **2011**, *124* (1), 109–127.

(189) Yang, C.; Tarkhov, A.; Rg Marusczyk, J.; Bienfait, B.; Gasteiger, J.; Kleinoeder, T.; Magdziarz, T.; Sacher, O.; Schwab, C. H.; Schwoebel, J.; et al. New Publicly Available Chemical Query Language, CSRML, To Support Chemotype Representations for Application to Data Mining and Modeling. *J. Chem. Inf. Model.* **2015**, *55* (3), 510-528.

(190) Russo, D. P.; Zorn, K. M.; Clark, A. M.; Zhu, H.; Ekins, S. Comparing Multiple Machine Learning Algorithms and Metrics for Estrogen Receptor Binding Prediction. *Mol. Pharm.* **2018**, *15* (10), 4361–4370.

(191) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016; pp 785–794.

(192) Jagielski, M.; Oprea, A.; Biggio, B.; Liu, C.; Nita-Rotaru, C.; Li, B. Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning. In 2018 IEEE Symposium on Security and Privacy (SP), 2018.

(193) Stulp, F.; Sigaud, O. Many regression algorithms, one unified model: A review. *Neural Networks* **2015**, *69*, 60-79. DOI: 10.1016/j.neunet.2015.05.005.

(194) De Menezes, L. M. B.; Volpato, M. C.; Rosalen, P. L.; Cury, J. A. Bone as a Biomarker of Acute Fluoride Toxicity. *Forensic Sci. Int.* **2003**, *137* (2–3), 209–214.

(195) Nandi, S.; Vracko, M.; Bagchi, M. C. Anticancer Activity of Selected Phenolic Compounds: QSAR Studies Using Ridge Regression and Neural Networks. *Chemical Biology & Drug Design* **2007**, *70* (5), 424-436. DOI: 10.1111/j.1747-0285.2007.00575.x.

(196) Das, R. N.; Roy, K. Development of classification and regression models for Vibrio fischeri toxicity of ionic liquids: green solvents for the future. *Toxicology Research* **2012**, *1* (3), 186-195. DOI: 10.1039/c2tx20020a.

(197) Borchert, D. M.; Walgenbach, J. F.; Kennedy, G. G.; Long, J. W. Toxicity and Residual Activity of Methoxyfenozide and Tebufenozide to Codling Moth (Lepidoptera: Tortricidae) and Oriental Fruit Moth (Lepidoptera: Tortricidae). *Journal of Economic Entomology* **2004**, *97* (4), 1342-1352. DOI: 10.1093/jee/97.4.1342.

(198) Tibshirani, R. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **2011**, *73* (3), 273-282. DOI: 10.1111/j.1467-9868.2011.00771.x.

(199) Micevska, T.; Warne, M. S. J.; Pablo, F.; Patra, R. Variation in, and Causes of, Toxicity of Cigarette Butts to a Cladoceran and Microtox. *Archives of Environmental Contamination and Toxicology* **2006**, *50* (2), 205-212. DOI: 10.1007/s00244-004-0132-y.

(200) Osborne, M. R.; Presnell, B.; Turlach, B. A. On the LASSO and its Dual. *Journal of Computational and Graphical Statistics* **2000**, *9* (2), 319-337. DOI: 10.1080/10618600.2000.10474883.

(201) Isbister, G. K.; O'Regan, L.; Sibbritt, D.; Whyte, I. M. Alprazolam is relatively more toxic than other benzodiazepines in overdose. *Br J Clin Pharmacol* **2004**, *58* (1), 88-95. DOI: 10.1111/j.1365-2125.2004.02089.x PubMed.

(202) Hawkins, D. M.; Basak, S. C.; Mills, D. QSARs for chemical mutagens from structure: ridge regression fitting and diagnostics. *Environmental Toxicology and Pharmacology* **2004**, *16* (1-2), 37-44. DOI: 10.1016/j.etap.2003.09.001.

(203) Loganayagam, A.; Arenas Hernandez, M.; Corrigan, A.; Fairbanks, L.; Lewis, C. M.; Harper, P.; Maisey, N.; Ross, P.; Sanderson, J. D.; Marinaki, A. M. Pharmacogenetic variants in the DPYD, TYMS, CDA and MTHFR genes are clinically significant predictors of fluoropyrimidine toxicity. *Br J Cancer* **2013**, *108* (12), 2505-2515. DOI: 10.1038/bjc.2013.262 PubMed.

(204) Cawley, G. C.; Talbot, N. L. Reduced rank kernel ridge regression. *Neural Processing Letters* **2002**, *16* (3), 293-302. DOI: 10.1023/a:1021798002258.

(205) Roy, K.; Ghosh, G. QSTR with extended topochemical atom (ETA) indices. 9. Comparative QSAR for the toxicity of diverse functional organic compounds to Chlorella vulgaris using chemometric tools. *Chemosphere* **2007**, *70* (1), 1-12. DOI: 10.1016/j.chemosphere.2007.07.037.

(206) Agarwal, V.; Gribok, A. V.; Koschan, A.; Abidi, M. A. Estimating Illumination Chromaticity via Kernel Regression. In 2006 International Conference on Image Processing, 2006.

(207) Kar, S.; Roy, K. QSAR modeling of toxicity of diverse organic chemicals to Daphnia magna using 2D and 3D descriptors. *Journal of Hazardous Materials* **2010**, *177* (1-3), 344-351. DOI: 10.1016/j.jhazmat.2009.12.038.

(208) Duchowicz, P. R. Linear Regression QSAR Models for Polo-Like Kinase-1 Inhibitors. *Cells* **2018**, *7* (2), 13. DOI: 10.3390/cells7020013 PubMed.

(209) Pan, Y.; Jiang, J.; Wang, R.; Cao, H. Advantages of support vector machine in QSPR studies for predicting auto-ignition temperatures of organic compounds. *Chemometrics and Intelligent Laboratory Systems* **2008**, *92* (2), 169-178. DOI: 10.1016/j.chemolab.2008.03.002.

(210) Razi, M.; Athappilly, K. A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. *Expert Systems with Applications* **2005**, *29* (1), 65-74. DOI: 10.1016/j.eswa.2005.01.006.

(211) Bermejo, S.; Cabestany, J. Learning with Nearest Neighbour Classifiers. *Neural Processing Letters* **2001**, *13*, 159-181. DOI: 10.1023/a:1011332406386.

(212) Neelamegam, S.; Ramaraj, E. Classification Algorithm in Data Mining: An Overview. *Int. J. P2P Netw. Trends Technol.* **2013**, *4* (8), 369–374.

(213) Samsudin, N. A.; Bradley, A. P. Nearest neighbour group-based classification. *Pattern Recognition* **2010**, *43* (10), 3458-3467. DOI: 10.1016/j.patcog.2010.05.010.

(214) Friel, N.; Pettitt, A. N. Classification using distance nearest neighbours. *Statistics and Computing* **2011**, *21* (3), 431-437. DOI: 10.1007/s11222-010-9179-y.

(215) Lakshmi, S. V.; Prabakaran, T. E. Application of K-Nearest Neighbour Classification Method for Intrusion Detection in Network Data. *International Journal of Computer Applications* **2014**, *97* (7), 34-37.

(216) Safavian, S. R.; Landgrebe, D. A Survey of Decision Tree Classifier Methodology. *IEEE Trans. Syst. Man. Cybern.* **1991**, *21* (3), 660–674.

(217) Song, Y. Y.; Lu, Y. Decision Tree Methods: Applications for Classification and Prediction. *Shanghai Arch. Psychiatry* **2015**, *27* (2), 130.

(218) Mahesh, B. Machine Learning Algorithms: A Review. *International Journal of Science and Research (IJSR)* **2020**, *9* (1), 381-386.

(219) Panda, M.; Patra, M. R. Network Intrusion Detection Using Naive Bayes. *International journal of computer science and network security* **2007**, *7* (12), 258-263.

(220) Martinez-Arroyo, M.; Sucar, L. E. Learning an Optimal Naive Bayes Classifier. In 18th International Conference on Pattern Recognition (ICPR'06), 2006.

(221) Taheri, S.; Mammadov, M. Learning the naive Bayes classifier with optimization models. *International Journal of Applied Mathematics and Computer Science* **2013**, *23* (4), 787-795. DOI: 10.2478/amcs-2013-0059.

(222) Wei, W.; Visweswaran, S.; Cooper, G. F. The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data. *J Am Med Inform Assoc* **2011**, *18* (4), 370-375. DOI: 10.1136/amiajnl-2011-000101 PubMed.

(223) Klon, A. E.; Lowrie, J. F.; Diller, D. J. Improved Naïve Bayesian Modeling of Numerical Data for Absorption, Distribution, Metabolism and Excretion (ADME) Property Prediction. *Journal of Chemical Information and Modeling* **2006**, *46* (5), 1945-1956. DOI: 10.1021/ci0601315.

(224) Frank, E.; Hall, M.; Pfahringer, B. Locally Weighted Naive Bayes. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, 2002; pp 249–256.

(225) Jiang, L.; Zhang, H.; Zhihua, C. A Novel Bayes Model: Hidden Naive Bayes. *IEEE Transactions on Knowledge and Data Engineering* **2008**, *21* (10), 1361-1371. DOI: 10.1109/tkde.2008.234.

(226) Cortes, C.; Vapnik, V. Support-vector networks. *Machine Learning* **1995**, *20* (3), 273-297. DOI: 10.1007/bf00994018.

(227) Vapnik, V. Principles of Risk Minimization for Learning Theory. In *Advances in neural information processing systems*, 1991; Vol. 4, pp 831–838.

(228) Zhong, M.; Nie, X.; Yan, A.; Yuan, Q. Carcinogenicity Prediction of Noncongeneric Chemicals by a Support Vector Machine. *Chemical Research in Toxicology* **2013**, *26* (5), 741-749. DOI: 10.1021/tx4000182.

(229) Burges, C. J. C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data mining and knowledge discovery* **1998**, *2* (2), 121-167.

(230) Vapnik, V. N. An overview of statistical learning theory. *IEEE Transactions on Neural Networks* **1999**, *10* (5), 988-999. DOI: 10.1109/72.788640.

(231) Ben-Hur, A.; Weston, J. A User's Guide to Support Vector Machines. In *Data Mining Techniques for the Life Sciences*, Carugo, O., Eisenhaber, F. Ed.; Vol. 609; Humana Press, 2010; pp 223-239.

(232) Noble, W. S. What is a support vector machine? *Nature Biotechnology* **2006**, *24* (12), 1565-1567. DOI: 10.1038/nbt1206-1565.

(233) Rebentrost, P.; Mohseni, M.; Lloyd, S. Quantum Support Vector Machine for Big Data Classification. *Physical Review Letters* **2014**, *113* (13), 130503. DOI: 10.1103/physrevlett.113.130503.

(234) Amari, S.; Wu, S. Improving support vector machine classifiers by modifying kernel functions. *Neural Networks* **1999**, *12* (6), 783-789. DOI: 10.1016/s0893-6080(99)00032-5.

(235) Varewyck, M.; Martens, J. P. A Practical Approach to Model Selection for Support Vector Machines with a Gaussian Kernel. *IEEE Trans. Syst. Man, Cybern. Part B Cybern.* **2010**, *41* (2), 330–340.

(236) Hussain, M.; Wajid, S. K.; Elzaart, A.; Berbar, M. A Comparison of SVM Kernel Functions for Breast Cancer Detection. In 2011 Eighth International Conference Computer Graphics, Imaging and Visualization, 2011.

(237) Kuo, B. C.; Ho, H. H.; Li, C. H.; Hung, C. C.; Taur, J. S. A Kernel-Based Feature Selection Method for SVM with RBF Kernel for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2013**, *7* (1), 317–326.

(238) Scholkopf, B.; Kah-Kay, S.; Burges, C. J. C.; Girosi, F.; Niyogi, P.; Poggio, T.; Vapnik, V. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Transactions on Signal Processing* **1997**, *45* (11), 2758-2765. DOI: 10.1109/78.650102.

(239) Xiao, Y.; Wang, H.; Zhang, L.; Xu, W. Two methods of selecting Gaussian kernel parameters for one-class SVM and their application to fault detection. *Knowledge-Based Systems* **2014**, *59*, 75-84. DOI: 10.1016/j.knosys.2014.01.020.

(240) Kim, H.-C.; Pang, S.; Je, H.-M.; Kim, D.; Yang Bang, S. Constructing support vector machine ensemble. *Pattern Recognition* **2003**, *36* (12), 2757-2767. DOI: 10.1016/s0031-3203(03)00175-4.

(241) Mangasarian, O. L.; Wild, E. W. Multisurface proximal support vector machine classification via generalized eigenvalues. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2005**, *28* (1), 69-74. DOI: 10.1109/tpami.2006.17.

(242) Khemchandani, R.; Chandra, S. Twin Support Vector Machines for Pattern Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29* (5), 905–910.

(243) Shao, Y. H.; Zhang, C. H.; Wang, X. B.; Deng, N. Y. Improvements on Twin Support Vector Machines. *IEEE Trans. Neural Networks* **2011**, *22* (6), 962–968.

(244) Dioşan, L.; Rogozan, A.; Pecuchet, J.-P. Improving classification performance of Support Vector Machine by genetically optimising kernel shape and hyper-parameters. *Applied Intelligence* **2012**, *36*, 280-294. DOI: 10.1007/s10489-010-0260-1.

(245) Lanckriet, G. R.; Bartlett, P.; El Ghaoui, L.; Jordan, M. I.; Cristianini, N.; Jordan Lanckriet, M. I.; Ghaoui, E. Learning the Kernel Matrix with Semidefinite Programming. *Journal of Machine learning research* **2004**, *5* (Jan), 27-72.

(246) Sonnenburg, S.; Rätsch, G.; Schäfer, C.; Schölkopf, B. Large scale multiple kernel learning. *Journal of Machine Learning Research* **2006**, *7*, 1531-1565.

(247) Bucak, S. S.; Jin, R.; Jain, A. K. Multiple Kernel Learning for Visual Object Recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2013**, *36* (7), 1354–1369.

(248) Sagi, O.; Rokach, L. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2018**, *8* (4), e1249. DOI: 10.1002/widm.1249.

(249) Rokach, L. Decision forest: Twenty years of research. *Information Fusion* **2016**, *27*, 111-125. DOI: 10.1016/j.inffus.2015.06.005.

(250) Gomes, H. M.; Barddal, J. P.; Enembreck, F.; Bifet, A. A Survey on Ensemble Learning for Data Stream Classification. *ACM Computing Surveys (CSUR)* **2017**, *50* (2), 1-36. DOI: 10.1145/3054925.

(251) Rokach, L. Ensemble-based classifiers. *Artificial Intelligence Review* **2010**, *33*, 1-39. DOI: 10.1007/s10462-009-9124-7.

(252) Ren, Y.; Zhang, L.; Suganthan, P. N. Ensemble Classification and Regression-Recent Developments, Applications and Future Directions [Review Article]. *IEEE Computational Intelligence Magazine* **2016**, *11* (1), 41-53. DOI: 10.1109/mci.2015.2471235.

(253) Woźniak, M.; Graña, M.; Corchado, E. A survey of multiple classifier systems as hybrid systems. *Information Fusion* **2014**, *16*, 3-17. DOI: 10.1016/j.inffus.2013.04.006.

(254) Krawczyk, B.; Minku, L. L.; Gama, J.; Stefanowski, J.; Woźniak, M. Ensemble learning for data stream analysis: A survey. *Information Fusion* **2017**, *37*, 132-156. DOI: 10.1016/j.inffus.2017.02.004.

(255) Galar, M.; Fernandez, A.; Barrenechea, E.; Bustince, H.; Herrera, F. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **2011**, *42* (4), 463-484. DOI: 10.1109/tsmcc.2011.2161285.

(256) Solomatine, D. P.; Shrestha, D. L. AdaBoost.RT: a boosting algorithm for regression problems. In 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541), 2004.

(257) Schapire, R. E. Explaining AdaBoost. In *Empirical Inference*, Schölkopf, B., Luo, Z., Vovk, V. Ed.; Springer, Berlin, Heidelberg, 2013; pp 37-52.

(258) Breiman, L. Bagging predictors. *Machine Learning* **1996**, *24* (2), 123-140. DOI: 10.1007/bf00058655.

(259) Schapire, R. E.; Freund, Y. A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting. *Journal of computer and system sciences* **1997**, *55* (1), 119-139.

(260) Dietterich, T. G. Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Mach. Learn.* **2000**, *40*, 139–157.

(261) Bernard, S.; Heutte, L.; Adam, S. On the selection of decision trees in Random Forests. In 2009 International Joint Conference on Neural Networks, 2009.

(262) Kingsford, C.; Salzberg, S. L. What are decision trees? *Nature biotechnology* **2008**, *26* (9), 1011-1013. DOI: 10.1038/nbt0908-1011 PubMed.

(263) Jiang, X.; Wu, C.-A.; Guo, H. Forest Pruning Based on Branch Importance. *Comput Intell Neurosci* **2017**, *2017*, 3162571. DOI: 10.1155/2017/3162571 PubMed.

(264) Yang, F.; Lu, W.-h.; Luo, L.-k.; Li, T. Margin optimization based pruning for random forest. *Neurocomputing* **2012**, *94*, 54-63. DOI: 10.1016/j.neucom.2012.04.007.

(265) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences* **2003**, *43* (6), 1947-1958. DOI: 10.1021/ci034160g.

(266) Ali, J.; Khan, R.; Ahmad, N.; Maqsood, I. Random Forests and Decision Trees. *Int. J. Comput. Sci. Issues* **2012**, *9* (5), 272–278.

(267) Prajwala, T. R. A Comparative Study on Decision Tree and Random Forest Using R Tool. *Int. J. Adv. Res. Comput. Commun. Eng.* **2015**, *4* (1), 196-199.

(268) Fawagreh, K.; Gaber, M. M.; Elyan, E. Random forests: from early developments to recent advancements. *Systems Science & Control Engineering* **2014**, *2* (1), 602-609. DOI: 10.1080/21642583.2014.956265.

(269) Kalogirou, S. A. Artificial neural networks in renewable energy systems applications: a review. *Renewable and Sustainable Energy Reviews* **2001**, *5* (4), 373-401. DOI: 10.1016/s1364-0321(01)00006-5.

(270) Markou, M.; Singh, S. Novelty detection: a review—part 2:: neural network based approaches. *Signal Processing* **2003**, *83* (12), 2499-2521. DOI: 10.1016/j.sigpro.2003.07.019.

(271) Egmont-Petersen, M.; de Ridder, D.; Handels, H. Image processing with neural networks—a review. *Pattern Recognition* **2002**, *35* (10), 2279-2301. DOI: 10.1016/s0031-3203(01)00178-9.

(272) Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks* **2015**, *61*, 85-117. DOI: 10.1016/j.neunet.2014.09.003.

(273) Greff, K.; Srivastava, R. K.; Koutnik, J.; Steunebrink, B. R.; Schmidhuber, J. LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems* **2016**, *28* (10), 2222-2232. DOI: 10.1109/tnnls.2016.2582924.

(274) Graves, A.; Mohamed, A.-r.; Hinton, G. Speech recognition with deep recurrent neural networks. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.

(275) Che, Z.; Purushotham, S.; Cho, K.; Sontag, D.; Liu, Y. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Sci Rep* **2018**, *8* (1), 6085. DOI: 10.1038/s41598-018-24271-9 PubMed.

(276) Saha, S.; Raghava, G. P. S. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins: Structure, Function, and Bioinformatics* **2006**, *65* (1), 40-48. DOI: 10.1002/prot.21078.

(277) Zhang, H.; Wang, Z.; Liu, D. A Comprehensive Review of Stability Analysis of Continuous-Time Recurrent Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems* **2014**, *25* (7), 1229-1262. DOI: 10.1109/tnnls.2014.2317880.

(278) Stathakis, D. How many hidden layers and nodes? *International Journal of Remote Sensing* **2009**, *30* (8), 2133-2147. DOI: 10.1080/01431160802549278.

(279) Teoh, E. J.; Tan, K. C.; Xiang, C. Estimating the Number of Hidden Neurons in a Feedforward Network Using the Singular Value Decomposition. *IEEE Transactions on Neural Networks* **2006**, *17* (6), 1623-1629. DOI: 10.1109/tnn.2006.880582.

(280) Verikas, A.; Bacauskiene, M. Feature selection with neural networks. *Pattern Recognition Letters* **2002**, *23* (11), 1323-1335. DOI: 10.1016/s0167-8655(02)00081-8.

(281) Sibi, P.; Jones, S. A.; Siddarth, P. Analysis of Different Activation Functions Using Back Propagation Neural Networks. *J. Theor. Appl. Inf. Technol.* **2013**, *47* (3), 1264–1268.

(282) Nair, V.; Hinton, G. E. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010; pp 807-814.

(283) Zeiler, M. D.; Ranzato, M.; Monga, R.; Mao, M.; Yang, K.; Le, Q. V.; Nguyen, P.; Senior, A.; Vanhoucke, V.; Dean, J.; et al. On rectified linear units for speech processing. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.

(284) Hara, K.; Saito, D.; Shouno, H. Analysis of function of rectified linear unit used in deep learning. In 2015 International Joint Conference on Neural Networks (IJCNN), 2015.

(285) Wray, J.; Green, G. G. Neural networks, approximation theory, and finite precision computation. *Neural Networks* **1995**, *8* (1), 31-37. DOI: 10.1016/0893-6080(94)00056-r.

(286) Siddique, M. N. H.; Tokhi, M. O. Training neural networks: backpropagation vs. genetic algorithms. In *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222)*, 2001; IEEE: Vol. 4, pp 2673–2678. DOI: 10.1109/ijcnn.2001.938792.

(287) Ismail Fawaz, H.; Forestier, G.; Weber, J.; Idoumghar, L.; Muller, P.-A. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery* **2019**, *33* (4), 917-963. DOI: 10.1007/s10618-019-00619-1.

(288) Wilamowski, B. M.; Hao, Y. Neural Network Learning Without Backpropagation. *IEEE Transactions on Neural Networks* **2010**, *21* (11), 1793-1803. DOI: 10.1109/tnn.2010.2073482.

(289) Zhu, X. X.; Tuia, D.; Mou, L.; Xia, G.-S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geoscience and Remote Sensing Magazine* **2017**, *5* (4), 8-36. DOI: 10.1109/mgrs.2017.2762307.

(290) Miotto, R.; Wang, F.; Wang, S.; Jiang, X.; Dudley, J. T. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics* **2018**, *19* (6), 1236-1246. DOI: 10.1093/bib/bbx044 PubMed.

(291) Guo, Y.; Liu, Y.; Oerlemans, A.; Lao, S.; Wu, S.; Lew, M. S. Deep learning for visual understanding: A review. *Neurocomputing* **2016**, *187*, 27-48. DOI: 10.1016/j.neucom.2015.09.116.

(292) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521* (7553), 436-444. DOI: 10.1038/nature14539.

(293) Kriegeskorte, N. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science* **2015**, *1*, 417-446. DOI: 10.1146/annurev-vision-082114-035447.

(294) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555* (7698), 604-610. DOI: 10.1038/nature25978.

(295) Esteva, A.; Kuprel, B.; Novoa, R. A.; Ko, J.; Swetter, S. M.; Blau, H. M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542* (7639), 115-118. DOI: 10.1038/nature21056.

(296) Sladojevic, S.; Arsenovic, M.; Anderla, A.; Culibrk, D.; Stefanovic, D. Deep Neural Networks Based Recognition of Plant Diseases by Leaf Image Classification. *Comput Intell Neurosci* **2016**, *2016*, 3289801. DOI: 10.1155/2016/3289801 PubMed.

(297) Cireşan, D. C.; Meier, U.; Gambardella, L. M.; Schmidhuber, J. Deep, Big, Simple Neural Nets for Handwritten Digit Recognition. *Neural Computation* **2010**, *22* (12), 3207-3220. DOI: 10.1162/neco_a_00052.

(298) Pallipuram, V. K.; Bhuiyan, M.; Smith, M. C. A comparative study of GPU programming models and architectures using neural networks. *The Journal of Supercomputing* **2012**, *61* (3), 673-718. DOI: 10.1007/s11227-011-0631-3.

(299) Huqqani, A. A.; Schikuta, E.; Ye, S.; Chen, P. Multicore and GPU Parallelization of Neural Networks for Face Recognition. *Procedia Computer Science* **2013**, *18*, 349-358. DOI: 10.1016/j.procs.2013.05.198.

(300) Oh, K.-S.; Jung, K. GPU implementation of neural networks. *Pattern Recognition* **2004**, *37* (6), 1311-1314. DOI: 10.1016/j.patcog.2004.01.013.

(301) Lin, S.-B. Limitations of shallow nets approximation. *Neural Networks* **2017**, *94*, 96-102. DOI: 10.1016/j.neunet.2017.06.016.

(302) Lawrence, S.; Giles, C. L.; Ah Chung, T.; Back, A. D. Face recognition: a convolutional neural-network approach. *IEEE Transactions on Neural Networks* **1997**, *8* (1), 98-113. DOI: 10.1109/72.554195.

(303) Rawat, W.; Wang, Z. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation* **2017**, *29* (9), 2352-2449. DOI: 10.1162/neco_a_00990.

(304) Wiatowski, T.; Bolcskei, H. A Mathematical Theory of Deep Convolutional Neural Networks for Feature Extraction. *IEEE Transactions on Information Theory* **2017**, *64* (3), 1845-1866. DOI: 10.1109/tit.2017.2776228.

(305) Lee, C.-Y.; Gallagher, P.; Tu, Z. Generalizing Pooling Functions in CNNs: Mixed, Gated, and Tree. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 2018/04/01, 2018; PMLR: pp 464-472. DOI: 10.1109/tpami.2017.2703082.

(306) Chami, I.; Ying, R.; Ré, C.; Leskovec, J. Hyperbolic Graph Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019; Vol. 32.

(307) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chem Sci* **2018**, *9* (2), 513-530. DOI: 10.1039/c7sc02664a PubMed.

(308) Niepert, M.; Ahmad, M.; Kutzkov, K. Learning Convolutional Neural Networks for Graphs. In *International Conference on Machine Learning*, 2016; PMLR: pp 2014-2023.

(309) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem Sci* **2018**, *10* (2), 370-377. DOI: 10.1039/c8sc04228d PubMed.

(310) Ying, R.; He, R.; Chen, K.; Eksombatchai, P.; Hamilton, W. L.; Leskovec, J. Graph Convolutional Neural Networks for Web-Scale Recommender Systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018/07/19, 2018; ACM: pp 974-983. DOI: 10.1145/3219819.3219890.

(311) Jimenez-Carretero, D.; Abrishami, V.; Fernández-de-Manuel, L.; Palacios, I.; Quílez-Álvarez, A.; Díez-Sánchez, A.; del Pozo, M. A.; Montoya, M. C. Tox_(R)CNN: Deep Learning-Based Nuclei Profiling Tool for Drug Toxicity Screening. *PLoS Comput. Biol.* **2018**, *14* (11), e1006238.

(312) Press, E. B.; Maron, D. M.; Ames, B. N. Revised Methods for the Salmonella Mutagenicity Test. *Mutation Research/Environmental Mutagenesis and Related Subjects* **1983**, *113* (3-4), 173-215.

(313) Ames, B. N.; Durston, W. E.; Yamasaki, E.; Lee, F. D. Carcinogens Are Mutagens: A Simple Test Combining Liver Homogenates for Activation and Bacteria for Detection. *Proc. Natl. Acad. Sci. U. S. A.* **1973**, *70* (8), 2281–2285.

(314) Hillebrecht, A.; Muster, W.; Brigo, A.; Kansy, M.; Weiser, T.; Singer, T. Comparative Evaluation of in Silico Systems for Ames Test Mutagenicity Prediction: Scope and Limitations. *Chem. Res. Toxicol.* **2011**, *24* (6), 843–854.

(315) Amberg, A.; Beilke, L.; Bercu, J.; Bower, D.; Brigo, A.; Cross, K. P.; Custer, L.; Dobo, K.; Dowdy, E.; Ford, K. A.; et al. Principles and procedures for implementation of ICH M7 recommended (Q) SAR analyses. *Regulatory Toxicology and Pharmacology* **2016**, *77*, 13-24.

(316) Novotarskyi, S.; Abdelaziz, A.; Sushko, Y.; Kö, R.; Vogt, J.; Tetko, I. V. ToxCast EPA in Vitro to in Vivo Challenge: Insight into the Rank-I Model. *Chem. Res. Toxicol.* **2016**, *29* (5), 768-775.

(317) Xu, T.; Ngan, D. K.; Ye, L.; Xia, M.; Xie, H. Q.; Zhao, B.; Simeonov, A.; Huang, R. Predictive Models for Human Organ Toxicity Based on In Vitro Bioactivity Data and Chemical Structure. *Chem. Res. Toxicol.* **2020**, *33* (3), 731–741.

(318) Rostami-Hodjegan, A. Physiologically Based Pharmacokinetics Joined with in Vitro-in Vivo Extrapolation of ADME: A Marriage under the Arch of Systems Pharmacology. *Clin. Pharmacol. Ther.* **2012**, *92* (1), 50–61.

(319) Caruana, R. Multitask Learning. *Mach. Learn.* **1997**, *28*, 41–75.

(320) Gibaja, E.; Ventura, S. Multi-Label Learning: A Review of the State of the Art and Ongoing Research. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2014**, *4* (6), 411–444.

(321) Sosnin, S.; Vashurina, M.; Withnall, M.; Karpov, P.; Fedorov, M.; Tetko, I. V. A Survey of Multi-Task Learning Methods in Chemoinformatics. *Molecular Informatics* **2019**, *38* (4), 1800108.

(322) Zhang, T. Analysis of Multi-Stage Convex Relaxation for Sparse Regularization. *J. Mach. Learn. Res.* **2010**, *11* (3), 1081–1107.

(323) Weigend, A. S.; Rumelhart, D. E.; Huberman, B. A. Generalization by weight-elimination applied to currency exchange rate prediction. In *[Proceedings] 1991 IEEE International Joint Conference on Neural Networks*, 1991; IEEE: pp 2374-2379. DOI: 10.1109/ijcnn.1991.170743.

(324) Hastie, T.; Rosset, S.; Tibshirani, R.; Zhu, J. The Entire Regularization Path for the Support Vector Machine. *J. Mach. Learn. Res.* **2004**, *5* (Oct), 1391–1415.

(325) Yao, Y.; Rosasco, L.; Caponnetto, A. On Early Stopping in Gradient Descent Learning. *Constructive Approximation* **2007**, *26* (2), 289-315. DOI: 10.1007/s00365-006-0663-2.

(326) Hagiwara, K. Regularization Learning, Early Stopping and Biased Estimator. *Neurocomputing* **2002**, *48* (1-4), 937–955.

(327) Ng, A. Y. Feature selection, L1 vs. L2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, 2004; ACM Press: p 78. DOI: 10.1145/1015330.1015435.

(328) Girosi, F.; Jones, M.; Poggio, T. Regularization Theory and Neural Networks Architectures. *Neural Computation* **1995**, *7* (2), 219-269. DOI: 10.1162/neco.1995.7.2.219.

(329) Williams, P. M. Bayesian Regularization and Pruning Using a Laplace Prior. *Neural Computation* **1995**, *7* (1), 117-143. DOI: 10.1162/neco.1995.7.1.117.

(330) Shi, G.; Zhang, J.; Li, H.; Wang, C. Enhance the Performance of Deep Neural Networks via L2 Regularization on the Input of Activations. *Neural Processing Letters* **2019**, *50*, 57-75. DOI: 10.1007/s11063-018-9883-8.

(331) Lee, S.-I.; Ganapathi, V.; Koller, D. Efficient Structure Learning of Markov Networks Using L 1-Regularization. In *Advances in neural Information processing systems*, 2006; Vol. 19, pp 817–824.

(332) Cronin, M. T.; Richarz, A. N.; Schultz, T. W. Identification and Description of the Uncertainty, Variability, Bias and Influence in Quantitative Structure-Activity Relationships (QSARs) for Toxicity Prediction. *Regul. Toxicol. Pharmacol.* **2019**, *106*, 90–104.

(333) Pittman, M. E.; Edwards, S. W.; Ives, C.; Mortensen, H. M. AOP-DB: A Database Resource for the Exploration of Adverse Outcome Pathways through Integrated Association Networks. *Toxicol. Appl. Pharmacol.* **2018**, *343*, 71–83.

(334) Leist, M.; Ghallab, A.; Graepel, R.; Marchan, R.; Hassan, R.; Bennekou, S. H.; Limonciel, A.; Vinken, M.; Schildknecht, S.; Waldmann, T.; et al. Adverse Outcome Pathways: Opportunities, Limitations and Open Questions. *Arch. Toxicol.* **2017**, *91*, 3477–3505.

(335) Vinken, M. The Adverse Outcome Pathway Concept: A Pragmatic Tool in Toxicology. *Toxicology* **2013**, *312*, 158–165.

(336) Allen, T. E. H.; Goodman, J. M.; Gutsell, S.; Russell, P. J. Defining Molecular Initiating Events in the Adverse Outcome Pathway Framework for Risk Assessment. *Chem. Res. Toxicol.* **2014**, *27* (12), 2100–2112.

(337) Ellison, C. M.; Enoch, S. J.; Cronin, M. T. D. A Review of the Use of in Silico Methods to Predict the Chemistry of Molecular Initiating Events Related to Drug Toxicity. *Expert Opin. Drug Metab. Toxicol.* **2011**, *7* (12), 1481–1495.

(338) Allen, T. E. H.; Goodman, J. M.; Gutsell, S.; Russell, P. J. A History of the Molecular Initiating Event. *Chemical Research in Toxicology* **2016**, *29* (12), 2060–2070.

(339) Slikker Jr, W.; de Souza Lima, T. A.; Archella, D.; de Silva Junior, J. B.; Barton-Maclaren, T.; Bo, L.; Buvinich, D.; Chaudhry, Q.; Chuan, P.; Deluyker, H.; et al. Emerging Technologies for Food and Drug Safety. *Regul. Toxicol. Pharmacol.* **2018**, *98*, 115–128.

(340) Zaunbrecher, V.; Beryt, E.; Parodi, D.; Telesca, D.; Doherty, J.; Malloy, T.; Allard, P. Has Toxicity Testing Moved into the 21st Century? A Survey and Analysis of Perceptions in the Field of Toxicology. *Environ. Health Perspect.* **2017**, *125* (8), 087024.

(341) Chesnut, M.; Yamada, T.; Adams, T.; Knight, D.; Kleinstreuer, N.; Kass, G.; Luechtefeld, T.; Hartung, T.; Maertens, A. Regulatory Acceptance of Read-Across. *ALTEX-Alternatives to Anim. Exp.* **2018**, *35* (3), 413–419.

(342) Mathea, M.; Klingspohn, W.; Baumann, K. Chemoinformatic Classification Methods and Their Applicability Domain. *Mol. Inform.* **2016**, *35* (5), 160–180.

(343) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminform.* **2015**, *7* (1), 1–13.

(344) Roy, K.; Kar, S.; Ambure, P. On a Simple Approach for Determining Applicability Domain of QSAR Models. *Chemom. Intell. Lab. Syst.* **2015**, *145*, 22–29.

(345) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M. D., J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; et al. QSAR Modeling: Where Have You Been? Where Are You Going To? *Journal of Medicinal Chemistry* **2014**, *57* (12), 4977–5010.

(346) Norinder, U.; Carlsson, L.; Boyer, S.; Eklund, M. Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination. *J. Chem. Inf. Model.* **2014**, *54* (6), 1596–1603.

(347) Alvarsson, J.; McShane, S. A.; Norinder, U.; Spjuth, O. Predicting With Confidence: Using Conformal Prediction in Drug Discovery. *J. Pharm. Sci.* **2021**, *110* (1), 42–49.

(348) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Cherkasov, A.; Li, J.; Gramatica, P.; Hansen, K.; Schroeter, T.; Müller, K. R.; et al. Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. *J. Chem. Inf. Model.* **2010**, *50* (12), 2094–2111.

(349) Tetko, I. V.; Bruneau, P.; Mewes, H.-W.; Rohrer, D. C.; Poda, G. I. Can We Estimate the Accuracy of ADME-Tox Predictions? *Drug Discov. Today* **2006**, *11* (15-16), 700-707.

(350) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Öberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical Assessment of QSAR Models of Environmental Toxicity against Tetrahymena Pyriformis: Focusing on Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.* **2008**, *48* (9), 1733–1746.

(351) Liu, R.; Wallqvist, A. Molecular Similarity-Based Domain Applicability Metric Efficiently Identifies Out-of-Domain Compounds. *J. Chem. Inf. Model.* **2018**, *59* (1), 181–189.

(352) Vishwakarma, G.; Sonpal, A.; Hachmann, J. Metrics for Benchmarking and Uncertainty Quantification: Quality, Applicability, and Best Practices for Machine Learning in Chemistry. *Trends in Chemistry* **2021**, *3* (2), 146–156.

(353) Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A Comprehensive Survey on Transfer Learning. *Proc. IEEE* **2020**, *109* (1), 43–76.

(354) Liu, R.; Wallqvist, A. Merging Applicability Domains for in Silico Assessment of Chemical Mutagenicity. *J. Chem. Inf. Model.* **2014**, *54* (3), 793–800.

(355) Mervin, L. H.; Afzal, A. M.; Drakakis, G.; Lewis, R.; Engkvist, O.; Bender, A. Target Prediction Utilising Negative Bioactivity Data Covering Large Chemical Space. *J. Cheminform.* **2015**, *7*, 1–16.

(356) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res* **2019**, *47* (D1), D930-D940.

(357) EPA, U. S. *ToxCast Database*. https://www.epa.gov/chemical-research/toxicity-forecasting (accessed 12 Aug 2024).

(358) Allen, T. E. H.; Wedlake, A. J.; Gelžinytė, E.; Gong, C.; Goodman, J. M.; Gutsell, S.; Russell, P. J. Neural Network Activation Similarity: A New Measure to Assist Decision Making in Chemical Toxicology. *Chem. Sci.* **2020**, *11* (28), 7335–7348.

(359) 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* **2019**, *33* (2), 388-401. DOI: 10.1021/acs.chemrestox.9b00325 From NLM Medline.

(360) *RDKit: Open-source cheminformatics*. http://www.rdkit.org (accessed 12 Aug 2024).

(361) *Google Colab*. https://colab.research.google.com/notebooks/welcome.ipynb# (accessed 12 Aug 2024).

(362) ElRafey, A.; Wojtusiak, J. Recent advances in scaling-down sampling methods in machine learning. *Wiley Interdisciplinary Reviews: Computational Statistics* **2017**, *9* (6), e1414.

(363) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. PubChem in 2021: New Data Content and Improved Web Interfaces. *Nucleic Acids Res.* **2021**, *49* (D1), D1388–D1395.

(364) 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 From NLM Medline.

(365) 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 From NLM Medline.

(366) 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 From NLM PubMed-not-MEDLINE.

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

(368) 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 From NLM Medline.

(369) Corvi, R.; Spielmann, H.; Hartung, T. Alternative Approaches for Carcinogenicity and Reproductive Toxicity. In *The History of Alternative Test Methods in Toxicology*, Balls, M., Combes, R., Worth, A. Eds.; Academic Press, Elsevier Inc., 2019; pp 209-217.

(370) Estevan, C.; Pamies, D.; Vilanova, E.; Sogorb, M. A. OECD Guidelines for In Vivo Testing of Reproductive Toxicity. In *Reproductive and Developmental Toxicology*, Gupta, R. C. Ed.; Academic Press, Elsevier Inc., 2017; pp 163-178.

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

(372) 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 From NLM Medline.

(373) 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.

(374) 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 From NLM Medline.

(375) 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 From NLM Medline.

(376) Kim, K. B.; Kwack, S. J.; Lee, J. Y.; Kacew, S.; Lee, B. M. Current opinion on risk assessment of cosmetics. *Journal of Toxicology and Environmental Health, Part B* **2021**, *24* (4), 137-161. DOI: 10.1080/10937404.2021.1907264 From NLM Medline.

(377) 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 From NLM Medline.

(378) Weiss, K.; Khoshgoftaar, T. M.; Wang, D. A survey of transfer learning. *Journal of Big data* **2016**, *3* (1), 1-40.

(379) Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference* **2000**, *90* (2), 227-244.
(380) Hoyer, P. B. Reproductive toxicology: current and future directions. *Biochemical pharmacology* **2001**, *62* (12), 1557-1564.
(381) Anderson, J. A.; Petrere, J. A.; Sakowski, R.; Fitzgerald, J. E.; de la Iglesia, F. A. Teratology study in rats with amsacrine, an antineoplastic agent. *Fundam Appl Toxicol* **1986**, *7* (2), 214-220. DOI: 10.1016/0272-0590(86)90150-8 From NLM.
(382) Beaudoin, A. R.; Fisher, D. L. An in vivo/in vitro evaluation of teratogenic action. *Teratology* **1981**, *23* (1), 57-61. DOI: 10.1002/tera.1420230108 From NLM.
(383) Cao, F.; Souders II, C. L.; Li, P.; Pang, S.; Liang, X.; Qiu, L.; Martyniuk, C. J. Developmental neurotoxicity of maneb: Notochord defects, mitochondrial dysfunction and hypoactivity in zebrafish (Danio rerio) embryos and larvae. *Ecotoxicol Environ Saf* **2019**, *170*, 227-237. DOI: 10.1016/j.ecoenv.2018.11.110 From NLM.
(384) Company, B.-M. S. *TEQUIN® (gatifloxacin) Tablets*. Food and Drug Administration, https://www.accessdata.fda.gov/drugsatfda_docs/label/2004/21061s023,024,21062s026,037lbl.pdf (accessed 12 Aug 2024).
(385) DailyMed. *ALFUZOSIN HYDROCHLORIDE tablet, extended release*. https://dailymed.nlm.nih.gov/dailymed/drugInfo.cfm?setid=8a677905-e66d-44f7-b564-6c561387ed03 (accessed 12 Aug 2024).
(386) DailyMed. *TERAZOSIN HYDROCHLORIDE*. https://dailymed.nlm.nih.gov/dailymed/fda/fdaDrugXsl.cfm?setid=68877160-c88e-4b4e-a61d-abd4dff96742&type=display (accessed 12 Aug 2024).
(387) Danielsson, C.; Brask, J.; Sköld, A. C.; Genead, R.; Andersson, A.; Andersson, U.; Stockling, K.; Pehrson, R.; Grinnemo, K. H.; Salari, S.; et al. Exploration of human, rat, and rabbit embryonic cardiomyocytes suggests K-channel block as a common teratogenic mechanism. *Cardiovascular research* **2013**, *97* (1), 23-32.
(388) Doherty, P. A.; Smith, R. P.; Ferm, V. H. Comparison of the teratogenic potential of two aliphatic nitriles in hamsters: succinonitrile and tetramethylsuccinonitrile. *Fundam Appl Toxicol* **1983**, *3* (1), 41-48. DOI: 10.1016/s0272-0590(83)80171-7 From NLM.
(389) Guidechem. *5-Phenyl-6-Ethyl-2,4-Diaminopyrimidine 27653-49-2 wiki*. https://www.guidechem.com/encyclopedia/5-phenyl-6-ethyl-2-4-diaminopy-dic118767.html (accessed 12 Aug 2024).

(390) Harada, T.; Kimura, E.; Hirata-Koizumi, M.; Hirose, A.; Kamata, E.; Ema, M. Reproductive and developmental toxicity screening study of 4-aminophenol in rats. *Drug Chem Toxicol* **2008**, *31* (4), 473-486. DOI: 10.1080/01480540802390627 From NLM.

(391) Harris, S. B.; Schardein, J. L.; Ulrich, C. E.; Ridlon, S. A. Inhalation developmental toxicity study of propylene oxide in Fischer 344 rats. *Toxicol Sci* **1989**, *13* (2), 323-331.

(392) Horvath, C.; Druga, A. Action of the phenothiazine derivative methophenazine on prenatal development in rats. *Teratology* **1975**, *11* (3), 325-329. DOI: 10.1002/tera.1420110312 From NLM.

(393) Iqbal, M. M.; Aneja, A.; Rahman, A.; Megna, J.; Freemont, W.; Shiplo, M.; Nihilani, N.; Lee, K. The potential risks of commonly prescribed antipsychotics: during pregnancy and lactation. *Psychiatry (Edgmont)* **2005**, *2* (8), 36-44. From NLM.

(394) Karlsson, M.; Danielsson, B. R.; Nilsson, M. F.; Danielsson, C.; Webster, W. S. New proposals for testing drugs with IKr-blocking activity to determine their teratogenic potential. *Current pharmaceutical design* **2007**, *13* (29), 2979-2988.

(395) Kitamura, A. [Observation of micrognathic development in mouse fetus induced by sulfadimethoxine]. *Kokubyo Gakkai Zasshi* **1998**, *65* (1), 25-41. DOI: 10.5357/koubyou.65.25 From NLM.

(396) Kocher, W. [Patterns of effects of triethylenemelamine (TEM) and their treatment-phase specifity during organogenesis of the chick embryo : I. Changes in the dimensions of the body, the outer form and surface structures]. *Wilhelm Roux Arch Entwickl Mech Org* **1969**, *162* (2), 161-196. DOI: 10.1007/bf00573538 From NLM.

(397) Li, D. K.; Zhou, Q. D.; Qin, X. B.; Sun, R. M.; Zhu, X. L.; Cheng, H. J.; Wang, C. S.; He, J. P.; Qian, C.; Z., X. S.; et al. An epidemiological study on the effect of N, N'-methylene-bis-(2-amino-1,3,4-thiadiazole) (MATDA) on outcomes of pregnancy. *Teratology* **1986**, *33* (3), 289-297. DOI: 10.1002/tera.1420330306 From NLM.

(398) Morash, M. G.; Soanes, K. H.; Achenbach, J. C.; Ellis, L. D. Assessing the morphological and behavioral toxicity of catechol using larval zebrafish. *Int J Mol Sci* **2022**, *23* (14), 7985.

(399) Nelson, B. K.; Brightwell, W. S.; MacKenzie-Taylor, D. R.; Khan, A.; Burg, J. R.; Weigel, W. W.; Goad, P. T. Teratogenicity of n-propanol and isopropanol administered at high inhalation concentrations to rats. *Food Chem Toxicol* **1988**, *26* (3), 247-254. DOI: 10.1016/0278-6915(88)90126-3 From NLM.

(400) Neuper, L.; Kummer, D.; Forstner, D.; Guettler, J.; Ghaffari-Tabrizi-Wizsy, N.; Fischer, C.; Juch, H.; Nonn, O.; Gauster, M. Candesartan Does Not Activate PPARγ and Its Target Genes in Early Gestation Trophoblasts. *Int J Mol Sci* **2022**, *23* (20), 12326.

(401) Nikolić-Kokić, A.; Tatalović, N.; Brkljačić, J.; Mijović, M.; Nestorović, V.; Mijušković, A.; Oreščanin-Dušić, Z.; Vidonja Uzelac, T.; Nikolić, M.; Spasić, S.; et al. Antipsychotic Drug-Mediated Adverse Effects on Rat Testicles May Be Caused by Altered Redox and Hormonal Homeostasis. *Int J Mol Sci* **2022**, *23* (22), 13698.

(402) Padberg, S.; Wacker, E.; Meister, R.; Panse, M.; Weber-Schoendorfer, C.; Oppermann, M.; Schaefer, C. Observational cohort study of pregnancy outcome after first-trimester exposure to fluoroquinolones. *Antimicrob Agents Chemother* **2014**, *58* (8), 4392-4398. DOI: 10.1128/aac.02413-14 From NLM.

(403) Paget, G. E.; Thorpe, E. A TERATOGENIC EFFECT OF A SULPHONAMIDE IN EXPERIMENTAL ANIMALS. *Br J Pharmacol Chemother* **1964**, *23* (2), 305-312. DOI: 10.1111/j.1476-5381.1964.tb01588.x From NLM.

(404) Park, H.; You, H. H.; Song, G. Multiple toxicity of propineb in developing zebrafish embryos: Neurotoxicity, vascular toxicity, and notochord defects in normal vertebrate development. *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology* **2021**, *243*, 108993.

(405) Pedersen, L.; Nørgaard, M.; Skriver, M. V.; Olsen, J.; Sørensen, H. T. Prenatal exposure to loratadine in children with hypospadias: a nested case-control study within the Danish National Birth Cohort. *Am J Ther* **2006**, *13* (4), 320-324. DOI: 10.1097/00045391-200607000-00008 From NLM.

(406) Roebuck, B. D.; Carpenter, S. J. Teratogenic effects of azaserine in the Syrian golden hamster. *Experientia* **1983**, *39* (3), 324-326. DOI: 10.1007/bf01955329 From NLM.

(407) Rumeau-Rouquette, C.; Goujard, J.; Huel, G. Possible teratogenic effect of phenothiazines in human beings. *Teratology* **1977**, *15* (1), 57-64. DOI: 10.1002/tera.1420150108 From NLM.

(408) RxReasoner. *Prazepam Pharmacology - Active Ingredient*. https://www.rxreasoner.com/substances/prazepam/pharmacology (accessed 12 Aug 2024).

(409) Schroeder, R. E.; Rajakumar, P. A.; Devaskar, S. U. Effect of streptozotocin-induced maternal diabetes on fetal rat brain glucose transporters. *Pediatr Res* **1997**, *41* (3), 346-352. DOI: 10.1203/00006450-199703000-00007 From NLM.

(410) Schultz, T. W.; Dumont, J. N.; Epler, R. G. The embryotoxic and osteolathyrogenic effects of semicarbazide. *Toxicology* **1985**, *36* (2-3), 183-198.

(411) Sethi, S. Clozapine in pregnancy. *Indian J Psychiatry* **2006**, *48* (3), 196-197. DOI: 10.4103/0019-5545.31586 From NLM.

(412) Shah, R. M.; Schuing, R.; Benkhaial, G.; Young, A. V.; Burdett, D. Genesis of hadacidin-induced cleft palate in hamster: morphogenesis, electron microscopy, and determination of DNA synthesis, cAMP, and enzyme acid phosphatase. *Am J Anat* **1991**, *192* (1), 55-68. DOI: 10.1002/aja.1001920107 From NLM.

(413) Takeno, S.; Nakagawa, M.; Sakai, T. Teratogenic effects of nitrazepam in rats. *Res Commun Chem Pathol Pharmacol* **1990**, *69* (1), 59-70. From NLM.

(414) Vainio, H.; Hemminki, K.; Elovaara, E. Toxicity of styrene and styrene oxide on chick embryos. *Toxicology* **1977**, *8* (3), 319-325. DOI: 10.1016/0300-483x(77)90079-8 From NLM.

(415) Leeuwen, V.; C.J., H., T.; Seinen, W. Aquatic toxicological aspects of dithiocarbamates and related compounds. IV. Teratogenicity and histopathology in rainbow trout (Salmo gairdneri). *Aquatic toxicology* **1986**, *9* (2-3), 147-159.

(416) Westhoff, J. H.; Steenbergen, P. J.; Thomas, L. S.; Heigwer, J.; Bruckner, T.; Cooper, L.; Tönshoff, B.; Hoffmann, G. F.; Gehrig, J. In vivo high-content screening in zebrafish for developmental nephrotoxicity of approved drugs. *Frontiers in Cell and Developmental Biology* **2020**, *8*, 583.

(417) Wiley, M. J.; Joneja, M. G. The teratogenic effects of beta-aminopropionitrile in hamsters. *Teratology* **1976**, *14* (1), 43-52. DOI: 10.1002/tera.1420140107 From NLM.

(418) Wilson, K. S.; Malfair Taylor, S. C. Raltitrexed: optimism and reality. *Expert Opinion on Drug Metabolism & Toxicology* **2009**, *5* (11), 1447-1454.

(419) Wormser, U.; Izrael, M.; Van der Zee, E. A.; Brodsky, B.; Yanai, J. A chick model for the mechanisms of mustard gas neurobehavioral teratogenicity. *Neurotoxicology and teratology* **2005**, *27* (1), 65-71.

(420) Yanai, J.; Pinkas, A.; Seidler, F. J.; Ryde, I. T.; Van der Zee, E. A.; Slotkin, T. A. Neurobehavioral teratogenicity of sarin in an avian model. *Neurotoxicol Teratol* **2009**, *31* (6), 406-412. DOI: 10.1016/j.ntt.2009.07.007 From NLM.

(421) Ying, W.; Jang, F. F.; Teng, C.; Tai-Zhen, H. Apoptosis may involve in prenatally heroin exposed neurobehavioral teratogenicity? *Med Hypotheses* **2009**, *73* (6), 976-977. DOI: 10.1016/j.mehy.2009.04.059 From NLM.

(422) York, R. G.; Randall, J. L.; Scott, W. J., Jr. Teratogenicity of paraxanthine (1,7-dimethylxanthine) in C57BL/6J mice. *Teratology* **1986**, *34* (3), 279-282. DOI: 10.1002/tera.1420340307 From NLM.

(423) Wegner, S.; Browne, P.; Dix, D. Identifying reference chemicals for thyroid bioactivity screening. *Reproductive Toxicology* **2016**, *65*, 402-413.

(424) Browne, P.; Kleinstreuer, N. C.; Ceger, P.; Deisenroth, C.; Baker, N.; Markey, K.; Thomas, R. S.; Judson, R. J.; Casey, W. Development of a curated Hershberger database. *Reproductive Toxicology* **2018**, *81*, 259-271.

(425) Kleinstreuer, N. C.; Ceger, P.; Watt, E. D.; Martin, M.; Houck, K.; Browne, P.; Thomas, R. S.; Casey, W. M.; Dix, D. J.; Allen, D.; et al. Development and validation of a computational model for androgen receptor activity. *Chemical research in toxicology* **2017**, *30* (4), 946-964.

(426) Pinto, C. L.; Markey, K.; Dix, D.; Browne, P. Identification of candidate reference chemicals for in vitro steroidogenesis assays. *Toxicology in Vitro* **2018**, *47*, 103-119.

(427) 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 From NLM Medline.

(428) Jamalpoor, A.; Hartvelt, S.; Dimopoulou, M.; Zwetsloot, T.; Brandsma, I.; Racz, P. I.; Osterlund, T.; Hendriks, G. A novel human stem cell-based biomarker assay for in vitro assessment of developmental toxicity. *Birth Defects Research* **2022**, *114* (19), 1210-1228.

(429) 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 From NLM Medline.

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

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

(432) Lourith, N.; Kanlayavattanakul, M. Natural surfactants used in cosmetics: glycolipids. *International journal of cosmetic science* **2009**, *31* (4), 255-261.

(433) Hougaard, K. S. Next generation reproductive and developmental Toxicology: crosstalk into the future. *Frontiers in Toxicology* **2021**, *3*, 652571.

(434) Chen, Y.; Qu, H.; Li, X.; Wang, H. Effects of amoxicillin exposure at different stages, doses and courses of pregnancy on adrenal development in fetal mice. *Food and Chemical Toxicology* **2023**, *175*, 113754.

(435) Ma, F.; Li, Y.; Yu, Y.; Li, Z.; Lin, L.; Chen, Q.; Xu, Q.; Pan, P.; Wang, Y.; Ge, R. S. Gestational exposure to tebuconazole affects the development of rat fetal Leydig cells. *Chemosphere* **2021**, *262*, 127792.

(436) Jia, Z. L.; Zhu, C. Y.; Rajendran, R. S.; Xia, Q.; Liu, K. C.; Zhang, Y. Impact of airborne total suspended particles (TSP) and fine particulate matter (PM2. 5)-induced developmental toxicity in zebrafish (Danio rerio) embryos. *Journal of Applied Toxicology* **2022**, *42* (10), 1585-1602.

(437) Wu, Q.; Yan, W.; Cheng, H.; Liu, C.; Hung, T. C.; Guo, X.; Li, G. Parental transfer of microcystin-LR induced transgenerational effects of developmental neurotoxicity in zebrafish offspring. *Environmental Pollution* **2017**, *231*, 471-478.

(438) Paquette, A. G.; Marsit, C. J. The developmental basis of epigenetic regulation of HTR2A and psychiatric outcomes. *Journal of Cellular Biochemistry* **2014**, *115* (12), 2065-2072.

(439) Rock, K. D.; St Armour, G.; Horman, B.; Phillips, A.; Ruis, M.; Stewart, A. K.; Jima, D.; Muddiman, D. C.; Stapleton, H. M.; Patisaul, H. B. Effects of prenatal exposure to a mixture of organophosphate flame retardants on placental gene expression and serotonergic innervation in the fetal rat brain. *Toxicol Sci* **2020**, *176* (1), 203-223.

(440) Cassina, M.; Salviati, L.; Di Gianantonio, E.; Clementi, M. Genetic susceptibility to teratogens: state of the art. *Reproductive toxicology* **2012**, *34* (2), 186-191.

(441) Guillotin, S.; Delcourt, N. Studying the Impact of Persistent Organic Pollutants Exposure on Human Health by Proteomic Analysis: A Systematic Review. *Int J Mol Sci* **2022**, *23* (22), 14271.

(442) Li, X.; Chen, L.; Zhou, H.; Wang, J.; Zhao, C.; Pang, X. PFOA regulate adenosine receptors and downstream concentration-response cAMP-PKA pathway revealed by integrated omics and molecular dynamics analyses. *Science of the Total Environment* **2022**, *803*, 149910.

(443) Li, R.; He, P.; Cui, J.; Staufenbiel, M.; Harada, N.; Shen, Y. Brain endogenous estrogen levels determine responses to estrogen replacement therapy via regulation of BACE1 and NEP in female Alzheimer's transgenic mice. *Molecular neurobiology* **2013**, *47*, 857-867.

(444) Nahálková, J. Finding New Ways How to Control BACE1. *The Journal of Membrane Biology* **2022**, *255* (2-3), 293-318.

(445) Ramírez, A. R.; Castro, M. A.; Angulo, C.; Ramió, L.; Rivera, M. M.; Torres, M.; Rigau, T.; Rodríguez-Gil, J. E.; Concha, I. I. The presence and function of dopamine type 2 receptors in boar sperm: a possible role for dopamine in viability, capacitation, and modulation of sperm motility. *Biology of Reproduction* **2009**, *80* (4), 753-761.

(446) Lawlor, M.; Zigo, M.; Kerns, K.; Cho, I. K.; Easley IV, C. A.; Sutovsky, P. Spermatozoan Metabolism as a Non-Traditional Model for the Study of Huntington's Disease. *Int J Mol Sci* **2022**, *23* (13), 7163.

(447) Ali, B. H.; Adham, S. A.; Al Balushi, K. A.; Shalaby, A.; Waly, M. I.; Manoj, P.; Beegam, S.; Yuvaraju, P.; Nemmar, A. Reproductive toxicity to male mice of nose only exposure to water-pipe smoke. *Cellular Physiology and Biochemistry* **2015**, *35* (1), 29-37.
(448) Jenardhanan, P.; Mathur, P. P. Kinases as targets for chemical modulators: Structural aspects and their role in spermatogenesis. *Spermatogenesis* **2014**, *4* (2), e979113.
(449) Ge, J. C.; Qian, Q.; Gao, Y. H.; Zhang, Y. F.; Li, Y. X.; Wang, X.; Fu, Y.; Ma, Y. M.; Wang, Q. Toxic effects of Tripterygium glycoside tablets on the reproductive system of male rats by metabolomics, cytotoxicity, and molecular docking. *Phytomedicine* **2023**, *114*, 154813.
(450) Wu, S.; Li, X.; Li, P.; Li, T.; Huang, G.; Sun, Q.; Dinnyés, A.; Shang, L.; Xu, W. Developing rat testicular organoid models for assessing the reproductive toxicity of antidepression drugs in vitro: Testicular organoids to assess antidepressants toxicity. *Acta Biochimica et Biophysica Sinica* **2022**, *54* (11), 1748.

Chapter 7: Appendices

7.1: Chapter 2 Appendices

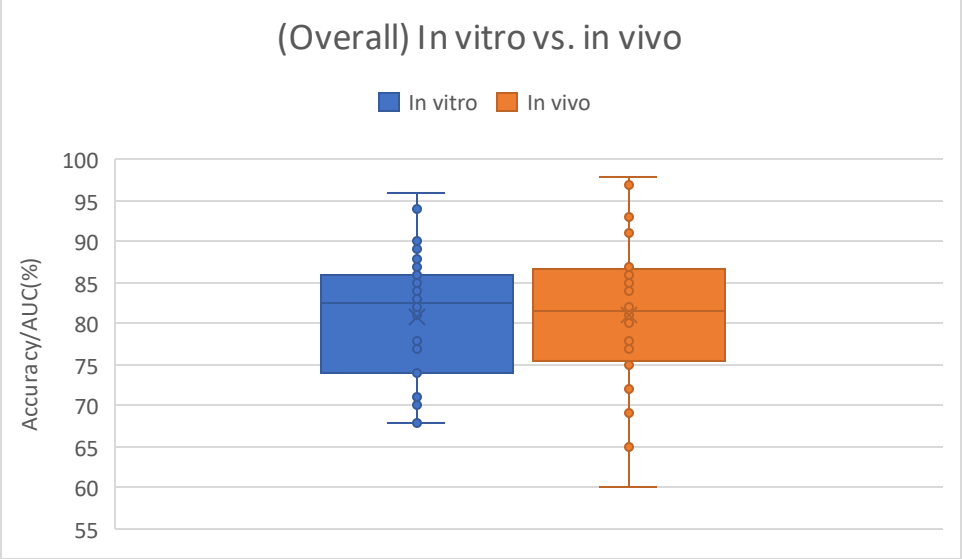


Figure A1: A plot of *in vitro* vs. *in vivo* results in Table 3 without considering toxicity endpoints

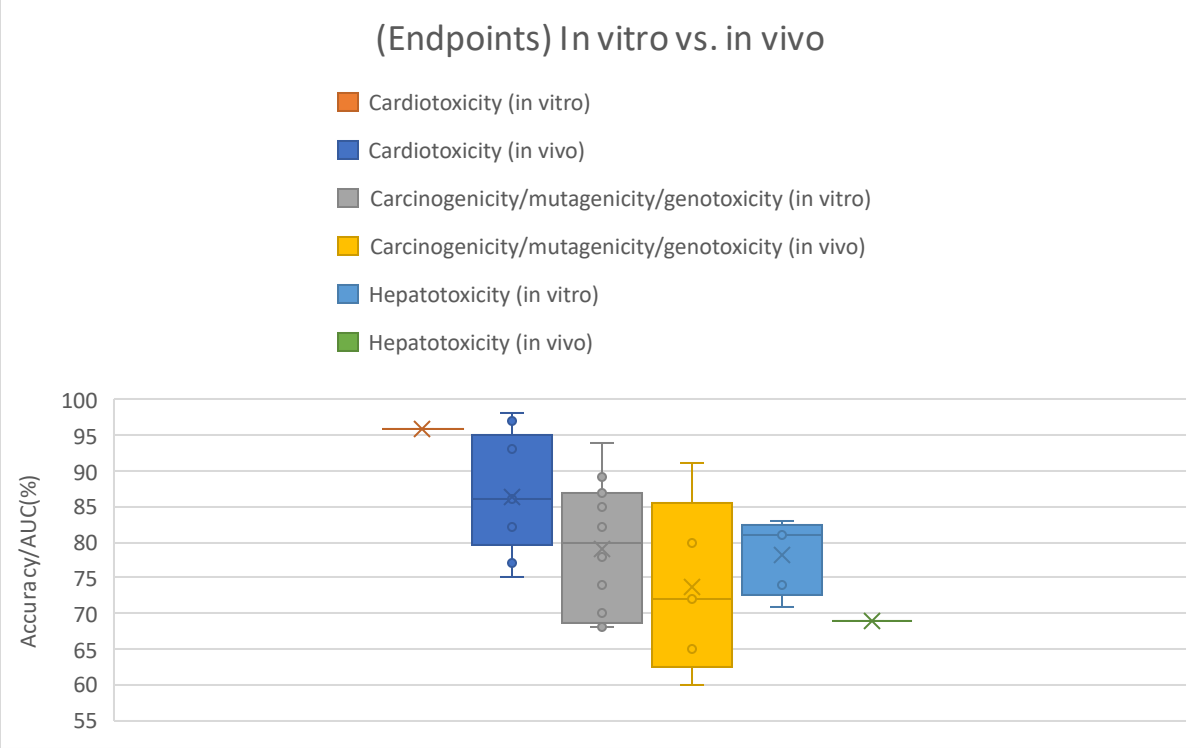


Figure A2: A plot of *in vitro* vs. *in vivo* results in Table 3 while considering toxicity endpoints

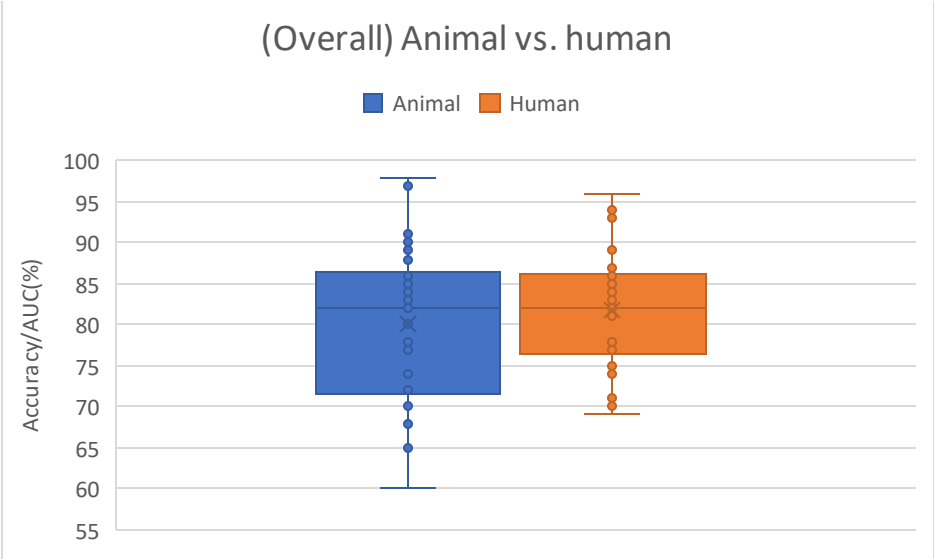


Figure A3: A plot of animal vs. human results in Table 3 without considering toxicity endpoints

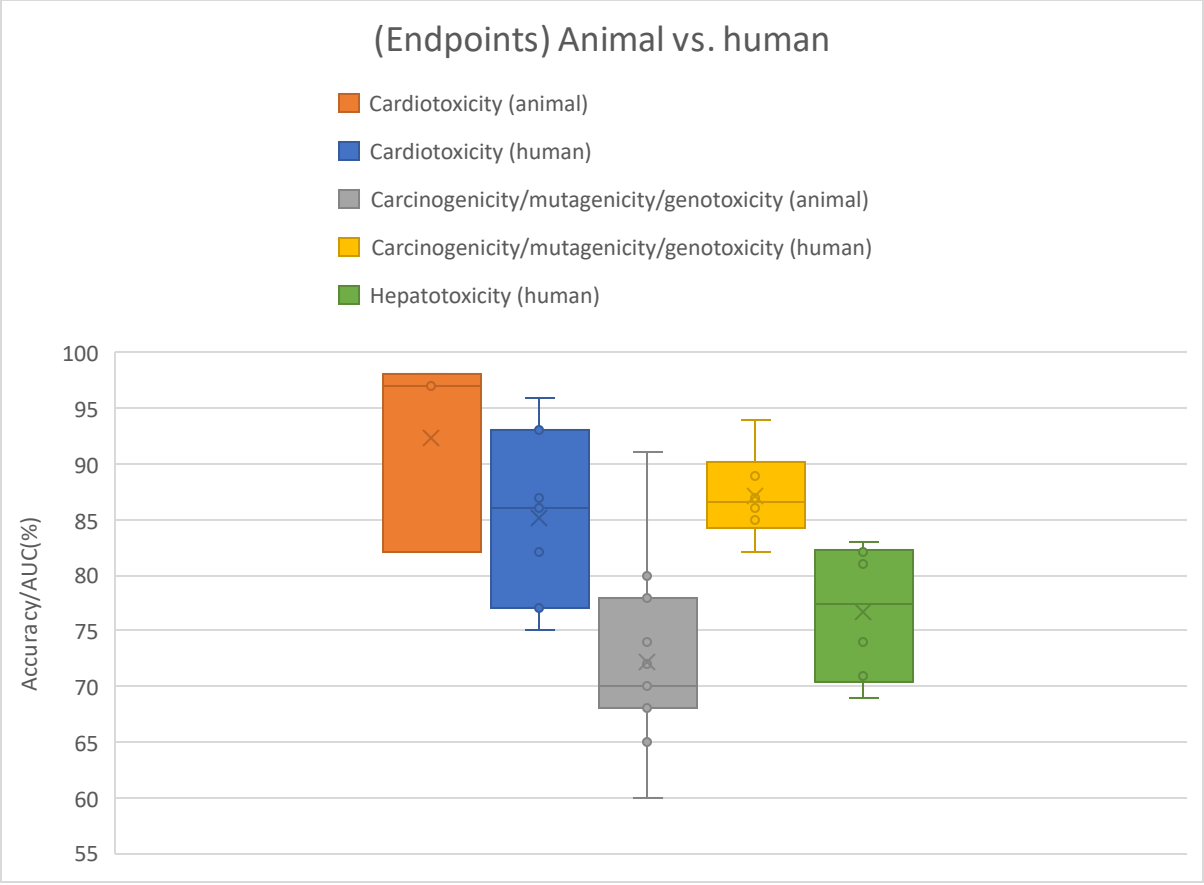


Figure A4: A plot of animal vs. human results in Table 3 without considering toxicity endpoints

7.2: Chapter 3 Appendices

Where applicable, the reported test accuracies as well as the judging of the better models for the tables and figures in this section were taken from the work by Allen et al[358]

Table A1: Absolute difference rates for the initial test of the similarity metric with varying thresholds and minimum number of molecules required for a molecule to be treated as similar to another. The reported accuracies used a [100,100] DNN architecture and were calculated using ECFP4 with a fingerprint length of 10000 in the study. A Morgan fingerprint length of 10000 was used to calculate the similarities.

Similarity, T	Min. no. of mol, N	Calculated training similarity, Q (%)	Calculated test similarity (%)	Average similarity, S (%)	Reported training accuracy (%)	Predicted test accuracy, P (%)	Reported test accuracy (%)	Absolute difference (DF) (%)	Target	Average absolute difference (average DF) (%)	Standard deviation for absolute difference (%)
0.1	1	98.4	96.8	97.6	91.2	89.0	84.7	4.3	AChE	4.3	6.4
		98.2	98.4	98.3	96	94.4	94.7	0.3	ADORA2A		
		97.2	98.2	97.7	92.6	90.5	90.1	0.4	AR		
		98.8	98.8	98.8	93.8	92.7	77.3	15.4	hERG		
		99	98.8	98.9	98.4	97.3	96	1.3	SERT		
	2	93.8	93.2	93.5	91.2	85.3	84.7	0.6	AChE	3.8	5.2
		96	97.2	96.6	96	92.7	94.7	2.0	ADORA2A		
		94	94	94	92.6	87.0	90.1	3.1	AR		
		97.2	95	96.1	93.8	90.1	77.3	12.8	hERG		
		97.2	96.8	97	98.4	95.4	96	0.6	SERT		
	3	88	89.2	88.6	91.2	80.8	84.7	3.9	AChE	4.5	1.9
		93.6	96.2	94.9	96	91.1	94.7	3.6	ADORA2A		
		92	91.4	91.7	92.6	84.9	90.1	5.2	AR		
		92.8	87.6	90.2	93.8	84.6	77.3	7.3	hERG		
		95.4	95	95.2	98.4	93.7	96	2.3	SERT		
0.15	1	97.4	95.2	96.3	91.2	87.8	84.7	3.1	AChE	5.2	5.8
		96.4	97.8	97.1	96	93.2	94.7	1.5	ADORA2A		
		90.6	90	90.3	92.6	83.6	90.1	6.5	AR		
		98.2	98	98.1	93.8	92.0	77.3	14.7	hERG		
		97.4	98	97.7	98.4	96.1	96	0.1	SERT		
	2	91	90.6	90.8	91.2	82.8	84.7	1.9	AChE	4.0	3.7
		93	96.2	94.6	96	90.8	94.7	3.9	ADORA2A		
		95.8	96.4	96.1	92.6	89.0	90.1	1.1	AR		
		94.4	92.4	93.4	93.8	87.6	77.3	10.3	hERG		
		94.8	94.8	94.8	98.4	93.3	96	2.7	SERT		
	3	84.6	85	84.8	91.2	77.3	84.7	7.4	AChE	6.7	2.5
		90.6	94.6	92.6	96	88.9	94.7	5.8	ADORA2A		
		85.8	85.6	85.7	92.6	79.4	90.1	10.7	AR		
		89	84.6	86.8	93.8	81.4	77.3	4.1	hERG		

		92.8	91.4	92.1	98.4	90.6	96	5.4	SERT		
0.2	1	93.4	89	91.2	91.2	83.2	84.7	1.5	AChE	4.8	2.6
		91.2	94.8	93	96	89.3	94.7	5.4	ADORA2A		
		89.2	87.8	88.5	92.6	82.0	90.1	8.1	AR		
		89.6	87.8	88.7	93.8	83.2	77.3	5.9	hERG		
		94.4	94.6	94.5	98.4	93.0	96	3.0	SERT		
	2	84	78.6	81.3	91.2	74.1	84.7	10.6	AChE	9.7	5.4
		86	91.8	88.9	96	85.3	94.7	9.4	ADORA2A		
		75.6	80.4	78	92.6	72.2	90.1	17.9	AR		
		79.6	79	79.3	93.8	74.4	77.3	2.9	hERG		
		90	89.6	89.8	98.4	88.4	96	7.6	SERT		
	3	74.2	71.6	72.9	91.2	66.5	84.7	18.2	AChE	15.9	5.7
		82	87.4	84.7	96	81.3	94.7	13.4	ADORA2A		
		68.2	73.2	70.7	92.6	65.5	90.1	24.6	AR		
		69.8	66.6	68.2	93.8	64.0	77.3	13.3	hERG		
		88.4	86.6	87.5	98.4	86.1	96	9.9	SERT		
0.25	1	85.8	78.6	82.2	91.2	75.0	84.7	9.7	AChE	12.2	4.5
		86.4	89.8	88.1	96	84.6	94.7	10.1	ADORA2A		
		75.8	76.4	76.1	92.6	70.5	90.1	19.6	AR		
		67.6	69.4	68.5	93.8	64.3	77.3	13.0	hERG		
		88.4	89.4	88.9	98.4	87.5	96	8.5	SERT		
	2	72.6	67.2	69.9	91.2	63.7	84.7	21.0	AChE	23.7	8.7
		78.2	82	80.1	96	76.9	94.7	17.8	ADORA2A		
		55.4	61.6	58.5	92.6	54.2	90.1	35.9	AR		
		52.2	50.4	51.3	93.8	48.1	77.3	29.2	hERG		
		83.6	81.6	82.6	98.4	81.3	96	14.7	SERT		
	3	57.6	59.6	58.6	91.2	53.4	84.7	31.3	AChE	32.6	11.4
		72.4	75.6	74	96	71.0	94.7	23.7	ADORA2A		
		43.4	52	47.7	92.6	44.2	90.1	45.9	AR		
		35.8	38.4	37.1	93.8	34.8	77.3	42.5	hERG		
		78	76.8	77.4	98.4	76.2	96	19.8	SERT		
0.3	1	76.2	71.8	74	91.2	67.5	84.7	17.2	AChE	22.2	9.5
		78.8	82.2	80.5	96	77.3	94.7	17.4	ADORA2A		
		55.8	58	56.9	92.6	52.7	90.1	37.4	AR		
		55.8	55.2	55.5	93.8	52.1	77.3	25.2	hERG		
		84.2	83.4	83.8	98.4	82.5	96	13.5	SERT		
	2	62	58	60	91.2	54.7	84.7	30.0	AChE	36.5	13.7
		68.6	70	69.3	96	66.5	94.7	28.2	ADORA2A		
		35.6	38.4	37	92.6	34.3	90.1	55.8	AR		
		33.6	34.2	33.9	93.8	31.8	77.3	45.5	hERG		
		74.2	74.2	74.2	98.4	73.0	96	23.0	SERT		
	3	47.6	45.2	46.4	91.2	42.3	84.7	42.4	AChE	45.9	15.1
		60	62.6	61.3	96	58.8	94.7	35.9	ADORA2A		
		23.8	27.2	25.5	92.6	23.6	90.1	66.5	AR		
		21	25.2	23.1	93.8	21.7	77.3	55.6	hERG		
		67.8	67.8	67.8	98.4	66.7	96	29.3	SERT		

Table A2: A comparison of the predicted values obtained using the similarity calculations with the published accuracies of the machine learning models for selected human biological targets. The predicted test accuracy was calculated by multiplying the training accuracy by the similarity between the datasets. A similarity threshold of 0.20 was used for the calculations in the algorithm and a fingerprint length of 10000 bits was used. Also, although the paper uses ECFP4, the algorithm uses Morgan fingerprints with a radius of 2, which has been shown to be comparable. The similarities were calculated using a sample size of 500 for both the training and test sets. The results from the paper uses a neural network architecture of [100,100].

No.	Target	Calculated similarity between training and test datasets (%)	Reported training accuracy (%)	Predicted test accuracy (%)	Reported test accuracy (%)	Absolute difference (%)
1	AChE	91.2	91.4	83.4	84.7	1.3
2	ADORA2A	93.8	96.6	90.6	94.7	4.1
3	ADRA2A	91.8	95.7	87.9	90.9	3.0
4	ADRB1	90.5	93.1	84.3	92.1	7.8
5	ADRB2	91.5	86.4	79.1	80.7	1.6
6	AGTR1	91.2	94.8	86.5	92.8	6.3
7	AKT1	92.1	96.1	88.5	93.7	5.2
8	AR	87.5	93.2	81.6	90.1	8.6
9	AVPR1A	91.7	98.4	90.3	96.1	5.8
10	BACE1	91.9	94.8	87.1	92.5	5.4
11	BCHE	91.4	92.4	84.5	86.3	1.8
12	CASP1	91.6	90.5	82.9	86.6	3.7
13	CASP3	93.9	92.6	87.0	89.1	2.1
14	CASP8	89.2	96.2	85.8	95.5	9.7
15	CHRM1	91.6	95	87.0	91.3	4.3
16	CHRM2	91.4	95.9	87.7	91.7	4.0
17	CHRM3	91.6	95.2	87.2	90.9	3.7
18	CHRM5	90.0	95.6	86.0	90.1	4.1
19	CHUK	91.1	95.8	87.3	92.9	5.6
20	CSF1R	89.5	97.9	87.6	95.3	7.7
21	CSNK1D	90.6	97.1	88.0	92.7	4.7
22	DRD1	89.9	92.5	83.2	88.2	5.0
23	DRD2	94.1	96.6	90.9	95.3	4.4
24	EDNRA	88.5	97.2	86.0	95.8	9.8
25	EDNRB	93.6	96.3	90.1	91.9	1.8

26	ELANE	92.5	95.0	87.9	91.6	3.7
27	EPHA2	89.5	96.9	86.7	94.9	8.2
28	FGFR1	91.6	96.0	87.9	93.5	5.6
29	FKBP1A	90.9	97.4	88.5	95.4	6.9
30	FLT1	91.3	97.8	89.3	95.6	6.3
31	FLT4	88.3	97.5	86.1	93.8	7.7
32	FYN	85.5	94.4	80.7	86.1	5.4
33	GSK3B	90.8	94.4	85.7	89.8	4.1
34	HDAC3	90.1	96.1	86.6	93.9	7.3
35	hERG	90.9	93.9	85.4	77.3	8.1
36	HRH1	90.7	95.4	86.5	91.7	5.2
37	HTR2A	93.7	97.7	91.5	96.1	4.6
38	HTR3A	89.9	95.3	85.7	92.4	6.7
39	IGF1R	92.2	97.1	89.5	94.5	5.0
40	INSR	91.3	96.9	88.5	93.9	5.4
41	KDR	91.8	93.2	85.6	91.6	6.0
42	LCK	93.6	94.7	88.6	91.8	3.2
43	LTB4R	89.3	96.6	86.3	95.0	8.7
44	LYN	88.8	96.4	85.6	93.1	7.5
45	MAPK1	93.1	79.9	74.4	78.3	3.9
46	MAPK9	92.3	97.7	90.2	93.5	3.3
47	MAPKAPK2	90.3	95.4	86.1	90.3	4.2
48	MET	90.7	97.3	88.3	94.7	6.4
49	MMP13	92.2	96.5	89.0	94.6	5.6
50	MMP2	92.3	95.4	88.1	91.5	3.4
51	MMP3	90.7	95.5	86.6	93.4	6.8
52	MMP9	91.4	89.3	81.6	85.9	4.3
53	NEK2	88.8	95.9	85.1	92.7	7.6
54	NR3C1	88.3	91.9	81.1	90.0	8.9
55	OPRD1	92.2	94.6	87.2	92.3	5.1
56	OPRM1	92.6	97.3	90.1	94.0	3.9
57	P2RY1	88.4	98.4	87.0	97.5	10.5
58	PAK4	89.0	98.0	87.2	94.1	6.9
59	PDE4A	89.1	95.9	85.5	91.5	6.0
60	PDE5A	91.1	95.2	86.7	91.1	4.4
61	PIK3CA	93.8	97.5	91.5	96.5	5.0
62	PPARG	89.3	88.5	79.0	86.4	7.4
63	PTPN1	92.5	92.1	85.2	80.8	4.4

64	PTPN11	87.7	91.7	80.4	84.2	3.8
65	PTPN2	90.1	92.9	83.7	87.6	3.9
66	RAF1	90.8	98.5	89.4	96.8	7.4
67	RARA	89.4	95.7	85.6	95.3	9.7
68	RARB	92.6	97.8	90.6	97.5	6.9
69	ROCK1	90.5	97.0	87.8	94.1	6.3
70	RPS6KA5	86.7	96.5	83.7	92.2	8.5
71	SERT	93.5	98.6	92.2	96.0	3.8
72	SIRT2	90.7	94.3	85.5	87.8	2.3
73	SIRT3	89.6	96.5	86.4	95.6	9.2
74	SLC6A2	91.8	95.7	87.9	93.2	5.3
75	SLC6A3	92.2	96.7	89.2	91.8	2.6
76	SRC	92.3	94.6	87.3	90.5	3.2
77	TACR2	91.5	96.2	88.0	95.0	7.0
78	TBXA2R	90.0	95.9	86.3	94.1	7.8
79	TEK	90.7	97.0	88.0	95.1	7.1

Table A3: A comparison of the predicted values obtained using the similarity calculations with the published accuracies of the machine learning models for selected human biological targets. The predicted test accuracy was calculated by multiplying the training accuracy by the similarity between the datasets. A similarity threshold of 0.20 was used for the calculations in the algorithm and a fingerprint length of 10000 bits was used. Also, although the paper uses ECFP4, the algorithm uses Morgan fingerprints with a radius of 2, which has been shown to be comparable. The similarities were calculated using a sample size of 500 for both the training and test sets. The results from the paper uses a neural network architecture of [1000,1000].

No.	Target	Calculated similarity between training and test datasets (%)	Reported training accuracy (%)	Predicted test accuracy (%)	Reported test accuracy (%)	Absolute difference (%)
1	AChE	91.2	95.8	87.4	84.4	3.0
2	ADORA2A	93.8	91.7	86.0	94.8	8.8
3	ADRA2A	91.8	97.2	89.3	91.2	1.9
4	ADRB1	90.5	94.9	85.9	92.1	6.2
5	ADRB2	91.5	85.4	78.1	80.4	2.3
6	AGTR1	91.2	96.6	88.1	93.5	5.4
7	AKT1	92.1	97.3	89.6	93.9	4.3
8	AR	87.5	92.5	80.9	89.2	8.3

9	AVPR1A	91.7	98.9	90.7	97.0	6.3
10	BACE1	91.9	95.4	87.7	92.6	4.9
11	BCHE	91.4	93.7	85.6	87.1	1.5
12	CASP1	91.6	88.8	81.3	85.9	4.6
13	CASP3	93.9	92.8	87.1	88.7	1.6
14	CASP8	89.2	96.9	86.4	96.0	9.6
15	CHRM1	91.6	96.4	88.3	91.9	3.6
16	CHRM2	91.4	97.0	88.7	92.4	3.7
17	CHRM3	91.6	96.7	88.6	91.2	2.6
18	CHRM5	90.0	96.3	86.6	90.2	3.6
19	CHUK	91.1	97.8	89.1	94.5	5.4
20	CSF1R	89.5	98.1	87.8	94.7	6.9
21	CSNK1D	90.6	94.4	85.5	90.8	5.3
22	DRD1	89.9	92.1	82.8	87.7	4.9
23	DRD2	94.1	97.7	91.9	95.6	3.7
24	EDNRA	88.5	97.6	86.4	95.8	9.4
25	EDNRB	93.6	97.6	91.4	93.9	2.5
26	ELANE	92.5	95.1	88.0	91.5	3.5
27	EPHA2	89.5	98.6	88.2	96.1	7.9
28	FGFR1	91.6	96.2	88.1	93.1	5.0
29	FKBP1A	90.9	96.2	87.4	95.7	8.3
30	FLT1	91.3	98.3	89.7	95.5	5.8
31	FLT4	88.3	98.3	86.8	93.4	6.6
32	FYN	85.5	95.4	81.6	88.2	6.6
33	GSK3B	90.8	96.1	87.3	89.9	2.6
34	HDAC3	90.1	96.8	87.2	93.8	6.6
35	hERG	90.9	92	83.6	76.2	7.4
36	HRH1	90.7	96.7	87.7	91.8	4.1
37	HTR2A	93.7	98.4	92.2	96.3	4.1
38	HTR3A	89.9	97.3	87.5	93.5	6.0
39	IGF1R	92.2	97.7	90.1	95.0	4.9
40	INSR	91.3	98.7	90.2	94.8	4.6
41	KDR	91.8	96.1	88.2	92.9	4.7
42	LCK	93.6	96.7	90.5	93.8	3.3
43	LTB4R	89.3	97.9	87.4	95.9	8.5
44	LYN	88.8	96.5	85.7	93.4	7.7
45	MAPK1	93.1	80.4	74.9	77.8	2.9
46	MAPK9	92.3	98.7	91.1	94.5	3.4

47	MAPKAPK2	90.3	95.8	86.5	90.2	3.7
48	MET	90.7	97.8	88.7	95.2	6.5
49	MMP13	92.2	97.2	89.6	94.7	5.1
50	MMP2	92.3	95.9	88.5	91.5	3.0
51	MMP3	90.7	95.9	87.0	93.1	6.1
52	MMP9	91.4	90.4	82.6	85.1	2.5
53	NEK2	88.8	97.2	86.3	92.8	6.5
54	NR3C1	88.3	91.9	81.1	89.8	8.7
55	OPRD1	92.2	96.1	88.6	92.3	3.7
56	OPRM1	92.6	96.9	89.7	93.6	3.9
57	P2RY1	88.4	98.5	87.1	97.3	10.2
58	PAK4	89.0	98.6	87.7	94.7	7.0
59	PDE4A	89.1	97.4	86.8	92.5	5.7
60	PDE5A	91.1	96.5	87.9	91.6	3.7
61	PIK3CA	93.8	98.2	92.1	96.9	4.8
62	PPARG	89.3	88.8	79.3	86.3	7.0
63	PTPN1	92.5	92.0	85.1	82.0	3.1
64	PTPN11	87.7	93.4	81.9	85.9	4.0
65	PTPN2	90.1	95.3	85.8	90.4	4.6
66	RAF1	90.8	99.0	89.9	96.5	6.6
67	RARA	89.4	97.3	87.0	96.0	9.0
68	RARB	92.6	99.2	91.9	98.6	6.7
69	ROCK1	90.5	97.9	88.6	93.8	5.2
70	RPS6KA5	86.7	98.0	85.0	93.5	8.5
71	SERT	93.5	99.1	92.7	96.6	3.9
72	SIRT2	90.7	95.1	86.3	89.6	3.3
73	SIRT3	89.6	96.4	86.3	92.0	5.7
74	SLC6A2	91.8	96.3	88.4	93.2	4.8
75	SLC6A3	92.2	95.5	88.1	91.4	3.3
76	SRC	92.3	96.6	89.2	91.1	1.9
77	TACR2	91.5	96.8	88.6	95.1	6.5
78	TBXA2R	90.0	96.4	86.8	94.0	7.2
79	TEK	90.7	98.1	89.0	95.0	6.0

Table A4: Results showing the best performing models for each target and the architecture used. Calculated absolute difference rates (%) for each architecture are also tabulated for each target. The better models were judged by the study.

		[100,100]	[1000,1000]	Better model judged by study		
No.	Target	Absolute difference (%)	Absolute difference (%)	[100,100]	[1000,1000]	Lower difference matches better model judged by study
1	AChE	1.3	3.0	1		1
2	ADORA2A	4.1	8.8		1	0
3	ADRA2A	3.0	1.9		1	1
4	ADRB1	7.8	6.2		1	1
5	ADRB2	1.6	2.3	1		1
6	AGTR1	6.3	5.4		1	1
7	AKT1	5.2	4.3		1	1
8	AR	8.6	8.3	1		0
9	AVPR1A	5.8	6.3		1	0
10	BACE1	5.4	4.9		1	1
11	BCHE	1.8	1.5		1	1
12	CASP1	3.7	4.6	1		1
13	CASP3	2.1	1.6	1		0
14	CASP8	9.7	9.6		1	1
15	CHRM1	4.3	3.6		1	1
16	CHRM2	4.0	3.7		1	1
17	CHRM3	3.7	2.6		1	1
18	CHRM5	4.1	3.6		1	1
19	CHUK	5.6	5.4		1	1
20	CSF1R	7.7	6.9	1		0
21	CSNK1D	4.7	5.3	1		1
22	DRD1	5.0	4.9	1		0
23	DRD2	4.4	3.7		1	1
24	EDNRA	9.8	9.4	1		0
25	EDNRB	1.8	2.5		1	0
26	ELANE	3.7	3.5	1		0
27	EPHA2	8.2	7.9		1	1
28	FGFR1	5.6	5.0	1		0
29	FKBP1A	6.9	8.3	1		1

30	FLT1	6.3	5.8	1		0
31	FLT4	7.7	6.6	1		0
32	FYN	5.4	6.6		1	0
33	GSK3B	4.1	2.6		1	1
34	HDAC3	7.3	6.6	1		0
35	hERG	8.1	7.4	1		0
36	HRH1	5.2	4.1		1	1
37	HTR2A	4.6	4.1		1	1
38	HTR3A	6.7	6.0		1	1
39	IGF1R	5.0	4.9		1	1
40	INSR	5.4	4.6		1	1
41	KDR	6.0	4.7		1	1
42	LCK	3.2	3.3		1	0
43	LTB4R	8.7	8.5		1	1
44	LYN	7.5	7.7		1	0
45	MAPK1	3.9	2.9	1		0
46	MAPK9	3.3	3.4		1	0
47	MAPKAPK2	4.2	3.7	1		0
48	MET	6.4	6.5		1	0
49	MMP13	5.6	5.1		1	1
50	MMP2	3.4	3.0		1	1
51	MMP3	6.8	6.1	1		0
52	MMP9	4.3	2.5	1		0
53	NEK2	7.6	6.5		1	1
54	NR3C1	8.9	8.7	1		0
55	OPRD1	5.1	3.7		1	1
56	OPRM1	3.9	3.9	1		0
57	P2RY1	10.5	10.2	1		0
58	PAK4	6.9	7.0		1	0
59	PDE4A	6.0	5.7		1	1
60	PDE5A	4.4	3.7		1	1
61	PIK3CA	5.0	4.8		1	1
62	PPARG	7.4	7.0	1		0
63	PTPN1	4.4	3.1		1	1
64	PTPN11	3.8	4.0		1	0
65	PTPN2	3.9	4.6		1	0
66	RAF1	7.4	6.6	1		0
67	RARA	9.7	9.0		1	1

68	RARB	6.9	6.7		1	1
69	ROCK1	6.3	5.2	1		0
70	RPS6KA5	8.5	8.5		1	1
71	SERT	3.8	3.9		1	0
72	SIRT2	2.3	3.3		1	0
73	SIRT3	9.2	5.7	1		0
74	SLC6A2	5.3	4.8		1	1
75	SLC6A3	2.6	3.3	1		1
76	SRC	3.2	1.9		1	1
77	TACR2	7.0	6.5		1	1
78	TBXA2R	7.8	7.2	1		0
79	TEK	7.1	6.0	1		0
Total				30	49	42
Percentage (%)				38.0	62.0	53.2

Table A5: A comparison of the predicted values obtained using the similarity calculations with the accuracies of the machine learning models for selected human biological targets. The predicted test accuracy was calculated by multiplying the training accuracy by the similarity between the datasets. A similarity threshold of 0.20 was used for the calculations in the algorithm and a fingerprint length of 10000 bits was used. The similarities were calculated using a sample size of 500 for both the training and test sets. The results were calculated using a neural network architecture of [100,100] and a model trained on the AChE training data. The test accuracies of the targets were obtained by using the AChE model to predict on the various targets in the table.

No.	Target	Calculated similarity between training and test datasets (%)	Model training accuracy (%)	Predicted test accuracy (%)	Model test accuracy (%)	Absolute difference (%)
1	ADORA2A	65.3	94.8	61.9	57.4	4.5
2	ADRA2A	74.2	94.8	70.3	73.2	2.9
3	ADRB1	70.2	94.8	66.5	61.2	5.4
4	ADRB2	73.8	94.8	69.9	62.1	7.8
5	AGTR1	69.5	94.8	65.9	65.0	0.9
6	AKT1	62.7	94.8	59.4	52.1	7.3
7	AR	61.1	94.8	57.9	73.6	15.7
8	AVPR1A	69.1	94.8	65.4	71.7	6.3
9	BACE1	61.3	94.8	58.1	46.6	11.5
10	BCHE	84.9	94.8	80.5	80.9	0.4
11	CASP1	70.4	94.8	66.7	58.2	8.5
12	CASP3	73.0	94.8	69.2	64.7	4.5

13	CASP8	55.4	94.8	52.5	70.3	17.8
14	CHRM1	72.6	94.8	68.8	59.3	9.5
15	CHRM2	70.4	94.8	66.7	71.2	4.5
16	CHRM3	75.0	94.8	71.1	66.7	4.4
17	CHRM5	75.9	94.8	71.9	72.4	0.5
18	CHUK	61.8	94.8	58.6	80.0	21.4
19	CSF1R	59.1	94.8	56.0	60.8	4.8
20	CSNK1D	64.4	94.8	61.1	65.7	4.6
21	DRD1	71.0	94.8	67.3	76.0	8.7
22	DRD2	81.1	94.8	76.9	66.1	10.8
23	EDNRA	63.7	94.8	60.4	53.2	7.1
24	EDNRB	60.6	94.8	57.4	58.6	1.2
25	ELANE	62.3	94.8	59.0	50.4	8.7
26	EPHA2	59.2	94.8	56.1	72.7	16.6
27	FGFR1	63.3	94.8	60.0	60.2	0.3
28	FKBP1A	67.0	94.8	63.5	85.1	21.5
29	FLT1	68.8	94.8	65.2	73.3	8.1
30	FLT4	68.1	94.8	64.5	70.9	6.4
31	FYN	59.0	94.8	55.9	78.1	22.2
32	GSK3B	67.7	94.8	64.2	52.9	11.3
33	HDAC3	70.8	94.8	67.1	55.7	11.4
34	hERG	70.9	94.8	67.2	55.2	12.0
35	HRH1	78.3	94.8	74.2	69.2	5.0
36	HTR2A	79.4	94.8	75.3	63.4	11.9
37	HTR3A	75.1	94.8	71.2	84.8	13.6
38	IGF1R	62.0	94.8	58.8	53.9	4.9
39	INSR	71.4	94.8	67.6	66.4	1.2
40	KDR	60.4	94.8	57.2	43.8	13.4
41	LCK	64.4	94.8	61.0	43.0	18.1
42	LTB4R	71.1	94.8	67.4	71.5	4.0
43	LYN	64.6	94.8	61.2	74.3	13.1
44	MAPK1	69.0	94.8	65.4	56.7	8.7
45	MAPK9	71.5	94.8	67.8	68.9	1.1
46	MAPKAPK2	66.8	94.8	63.3	70.6	7.3
47	MET	64.7	94.8	61.3	54.0	7.3
48	MMP13	70.7	94.8	67.0	45.7	21.3
49	MMP2	76.4	94.8	72.4	39.8	32.6
50	MMP3	75.3	94.8	71.4	43.8	27.5
51	MMP9	71.6	94.8	67.9	46.1	21.8
52	NEK2	65.5	94.8	62.1	81.7	19.6
53	NR3C1	65.8	94.8	62.4	71.1	8.8
54	OPRD1	77.5	94.8	73.5	66.2	7.2
55	OPRM1	76.3	94.8	72.3	65.4	6.9
56	P2RY1	57.8	94.8	54.8	72.6	17.8
57	PAK4	57.3	94.8	54.3	82.0	27.7

58	PDE4A	65.4	94.8	62.0	67.6	5.6
59	PDE5A	69.4	94.8	65.8	63.2	2.6
60	PIK3CA	52.0	94.8	49.3	42.1	7.2
61	PPARG	73.5	94.8	69.7	60.9	8.8
62	PTPN1	69.0	94.8	65.4	62.2	3.2
63	PTPN11	66.7	94.8	63.2	73.2	10.1
64	PTPN2	65.3	94.8	61.8	75.4	13.5
65	RAF1	61.3	94.8	58.1	55.9	2.2
66	RARA	65.2	94.8	61.8	83.9	22.1
67	RARB	63.7	94.8	60.4	87.6	27.2
68	ROCK1	66.2	94.8	62.7	61.4	1.4
69	RPS6KA5	60.0	94.8	56.9	82.5	25.6
70	SERT	71.3	94.8	67.6	54.5	13.0
71	SIRT2	68.7	94.8	65.1	80.5	15.4
72	SIRT3	65.2	94.8	61.7	86.6	24.9
73	SLC6A2	70.1	94.8	66.4	54.4	12.1
74	SLC6A3	70.0	94.8	66.3	59.6	6.8
75	SRC	62.2	94.8	59.0	57.5	1.5
76	TACR2	69.4	94.8	65.8	75.2	9.4
77	TBXA2R	67.5	94.8	64.0	69.3	5.3
78	TEK	60.8	94.8	57.6	70.6	13.0

Table A6: A comparison of the predicted values obtained using the similarity calculations with the accuracies of the machine learning models for selected human biological targets. The predicted test accuracy was calculated by multiplying the training accuracy by the similarity between the datasets. A similarity threshold of 0.20 was used for the calculations in the algorithm and a fingerprint length of 10000 bits was used. The similarities were calculated using a sample size of 500 for both the training and test sets. The results were calculated using a neural network architecture of [100,100] and a model trained on the ADORA2A training data. The test accuracies of the targets were obtained by using the ADORA2A model to predict on the various targets in the table.

No.	Target	Calculated similarity between training and test datasets (%)	Model training accuracy (%)	Predicted test accuracy (%)	Model test accuracy (%)	Absolute difference (%)
1	AChE	66.0	97.4	64.3	55.9	8.4
2	ADRA2A	70.0	97.4	68.2	62.6	5.6
3	ADRB1	63.5	97.4	61.9	58.4	3.5
4	ADRB2	71.0	97.4	69.2	58.3	10.9
5	AGTR1	73.0	97.4	71.1	86.2	15.1
6	AKT1	69.5	97.4	67.7	82.4	14.6
7	AR	60.9	97.4	59.3	73.4	14.0
8	AVPR1A	62.1	97.4	60.5	65.7	5.2

9	BACE1	54.3	97.4	52.9	50.3	2.6
10	BCHE	68.7	97.4	66.9	71.6	4.6
11	CASP1	65.2	97.4	63.5	57.3	6.2
12	CASP3	69.6	97.4	67.8	58.2	9.6
13	CASP8	55.8	97.4	54.3	76.3	22.0
14	CHRM1	66.1	97.4	64.4	41.9	22.5
15	CHRM2	61.4	97.4	59.8	56.1	3.8
16	CHRM3	65.3	97.4	63.6	51.1	12.5
17	CHRM5	62.5	97.4	60.9	63.3	2.4
18	CHUK	70.9	97.4	69.1	90.4	21.3
19	CSF1R	73.2	97.4	71.3	84.8	13.5
20	CSNK1D	78.0	97.4	76.0	91.5	15.5
21	DRD1	69.2	97.4	67.4	63.9	3.6
22	DRD2	69.7	97.4	67.9	35.6	32.3
23	EDNRA	61.4	97.4	59.8	59.3	0.6
24	EDNRB	59.4	97.4	57.9	68.0	10.2
25	ELANE	59.1	97.4	57.6	46.9	10.7
26	EPHA2	65.6	97.4	63.9	90.8	26.9
27	FGFR1	71.2	97.4	69.4	88.3	19.0
28	FKBP1A	59.9	97.4	58.4	70.1	11.8
29	FLT1	74.7	97.4	72.8	88.0	15.2
30	FLT4	75.0	97.4	73.0	84.3	11.3
31	FYN	71.3	97.4	69.5	90.0	20.5
32	GSK3B	80.1	97.4	78.1	72.6	5.5
33	HDAC3	69.8	97.4	68.0	67.7	0.3
34	hERG	64.0	97.4	62.3	49.2	13.2
35	HRH1	72.2	97.4	70.4	53.1	17.3
36	HTR2A	68.0	97.4	66.3	40.4	25.9
37	HTR3A	65.5	97.4	63.8	81.9	18.1
38	IGF1R	72.9	97.4	71.0	91.3	20.3
39	INSR	71.0	97.4	69.1	87.7	18.5
40	KDR	66.4	97.4	64.7	74.4	9.7
41	LCK	74.4	97.4	72.5	80.3	7.8
42	LTB4R	66.4	97.4	64.7	73.5	8.8
43	LYN	71.2	97.4	69.4	92.6	23.2
44	MAPK1	67.8	97.4	66.1	62.3	3.7
45	MAPK9	82.9	97.4	80.8	90.8	10.0
46	MAPKAPK2	72.2	97.4	70.4	83.4	13.0
47	MET	73.1	97.4	71.2	78.1	6.9
48	MMP13	70.5	97.4	68.7	44.3	24.4
49	MMP2	61.9	97.4	60.3	35.1	25.2
50	MMP3	59.9	97.4	58.4	40.9	17.5
51	MMP9	62.9	97.4	61.3	43.9	17.4
52	NEK2	74.3	97.4	72.4	91.1	18.7
53	NR3C1	56.3	97.4	54.9	72.6	17.7

54	OPRD1	58.1	97.4	56.6	38.8	17.9
55	OPRM1	59.1	97.4	57.6	44.9	12.7
56	P2RY1	66.7	97.4	65.0	81.5	16.6
57	PAK4	67.1	97.4	65.4	92.1	26.7
58	PDE4A	69.5	97.4	67.8	76.7	8.9
59	PDE5A	80.2	97.4	78.2	81.0	2.8
60	PIK3CA	72.3	97.4	70.5	82.5	12.0
61	PPARG	67.9	97.4	66.2	65.0	1.1
62	PTPN1	62.5	97.4	60.9	64.5	3.6
63	PTPN11	68.1	97.4	66.3	77.1	10.8
64	PTPN2	56.1	97.4	54.6	77.0	22.3
65	RAF1	66.9	97.4	65.2	77.3	12.1
66	RARA	59.5	97.4	58.0	87.1	29.2
67	RARB	56.7	97.4	55.3	90.3	35.0
68	ROCK1	75.1	97.4	73.2	80.2	7.0
69	RPS6KA5	64.0	97.4	62.3	92.2	29.8
70	SERT	60.2	97.4	58.7	31.5	27.1
71	SIRT2	67.1	97.4	65.3	77.3	11.9
72	SIRT3	59.9	97.4	58.4	89.6	31.2
73	SLC6A2	55.1	97.4	53.7	44.0	9.7
74	SLC6A3	54.3	97.4	52.9	44.0	8.9
75	SRC	76.9	97.4	74.9	80.8	5.9
76	TACR2	62.9	97.4	61.3	75.6	14.3
77	TBXA2R	63.9	97.4	62.3	69.4	7.2
78	TEK	69.5	97.4	67.7	82.8	15.0

Table A7: A comparison of the predicted values obtained using the similarity calculations with the accuracies of the machine learning models for selected human biological targets. The predicted test accuracy was calculated by multiplying the training accuracy by the similarity between the datasets. A similarity threshold of 0.20 was used for the calculations in the algorithm and a fingerprint length of 10000 bits was used. The similarities were calculated using a sample size of 500 for both the training and test sets. The results were calculated using a neural network architecture of [100,100] and a model trained on the AR training data. The test accuracies of the targets were obtained by using the AR model to predict on the various targets in the table.

No.	Target	Calculated similarity between training and test datasets (%)	Model training accuracy (%)	Predicted test accuracy (%)	Model test accuracy (%)	Absolute difference (%)
1	AChE	62.2	93.2	58.0	49.9	8.0
2	ADORA2A	54.8	93.2	51.1	37.1	13.9
3	ADRA2A	68.3	93.2	63.6	53.3	10.4

4	ADRB1	65.1	93.2	60.7	48.2	12.5
5	ADRB2	71.0	93.2	66.2	50.2	15.9
6	AGTR1	66.4	93.2	61.9	67.5	5.7
7	AKT1	53.7	93.2	50.0	44.5	5.5
8	AVPR1A	73.3	93.2	68.3	66.2	2.0
9	BACE1	53.8	93.2	50.1	38.5	11.6
10	BCHE	70.2	93.2	65.4	61.1	4.3
11	CASP1	65.6	93.2	61.1	66.0	4.9
12	CASP3	65.9	93.2	61.4	59.9	1.5
13	CASP8	68.0	93.2	63.3	75.2	11.9
14	CHRM1	62.1	93.2	57.9	37.2	20.7
15	CHRM2	65.5	93.2	61.0	55.1	5.9
16	CHRM3	63.6	93.2	59.3	42.8	16.5
17	CHRM5	69.8	93.2	65.1	62.7	2.4
18	CHUK	70.1	93.2	65.3	74.4	9.1
19	CSF1R	60.7	93.2	56.6	49.7	6.9
20	CSNK1D	67.4	93.2	62.8	62.0	0.8
21	DRD1	72.1	93.2	67.2	61.2	6.0
22	DRD2	56.6	93.2	52.7	17.3	35.4
23	EDNRA	61.4	93.2	57.2	49.4	7.9
24	EDNRB	64.0	93.2	59.6	58.8	0.8
25	ELANE	71.5	93.2	66.6	61.8	4.9
26	EPHA2	68.9	93.2	64.2	64.7	0.5
27	FGFR1	53.2	93.2	49.6	47.2	2.3
28	FKBP1A	67.2	93.2	62.6	69.0	6.4
29	FLT1	68.7	93.2	64.0	66.3	2.3
30	FLT4	66.6	93.2	62.1	61.4	0.7
31	FYN	72.7	93.2	67.7	75.1	7.4
32	GSK3B	60.3	93.2	56.2	41.7	14.4
33	HDAC3	70.9	93.2	66.1	52.8	13.3
34	hERG	54.8	93.2	51.1	43.3	7.8
35	HRH1	70.7	93.2	65.9	48.2	17.6
36	HTR2A	60.4	93.2	56.3	24.8	31.5
37	HTR3A	70.6	93.2	65.8	68.8	3.0
38	IGF1R	50.9	93.2	47.4	50.4	3.0
39	INSR	60.8	93.2	56.7	66.4	9.7
40	KDR	47.2	93.2	44.0	30.8	13.2
41	LCK	45.9	93.2	42.8	35.0	7.7
42	LTB4R	72.3	93.2	67.3	68.5	1.2
43	LYN	69.5	93.2	64.8	73.5	8.7
44	MAPK1	65.8	93.2	61.3	60.7	0.6
45	MAPK9	67.4	93.2	62.8	53.6	9.2
46	MAPKAPK2	64.0	93.2	59.6	66.2	6.6
47	MET	61.1	93.2	56.9	42.6	14.3
48	MMP13	61.8	93.2	57.6	36.4	21.2

49	MMP2	59.9	93.2	55.8	37.7	18.1
50	MMP3	57.7	93.2	53.8	37.6	16.1
51	MMP9	70.9	93.2	66.1	43.7	22.3
52	NEK2	75.2	93.2	70.1	79.9	9.8
53	NR3C1	84.6	93.2	78.8	83.3	4.5
54	OPRD1	56.5	93.2	52.6	30.4	22.2
55	OPRM1	61.4	93.2	57.2	39.0	18.2
56	P2RY1	67.5	93.2	62.9	66.4	3.5
57	PAK4	69.3	93.2	64.6	76.7	12.1
58	PDE4A	66.7	93.2	62.2	58.5	3.7
59	PDE5A	64.4	93.2	60.0	47.1	12.9
60	PIK3CA	62.6	93.2	58.3	46.7	11.6
61	PPARG	75.1	93.2	70.0	63.6	6.4
62	PTPN1	70.4	93.2	65.6	63.2	2.4
63	PTPN11	74.9	93.2	69.8	76.1	6.3
64	PTPN2	74.3	93.2	69.2	77.5	8.3
65	RAF1	68.5	93.2	63.8	72.5	8.6
66	RARA	80.0	93.2	74.5	86.9	12.4
67	RARB	80.6	93.2	75.1	89.6	14.5
68	ROCK1	65.7	93.2	61.2	52.9	8.3
69	RPS6KA5	73.6	93.2	68.5	80.0	11.5
70	SERT	55.7	93.2	51.9	23.5	28.4
71	SIRT2	73.6	93.2	68.5	74.8	6.2
72	SIRT3	72.1	93.2	67.2	82.1	14.9
73	SLC6A2	67.8	93.2	63.2	42.2	21.0
74	SLC6A3	66.4	93.2	61.9	42.9	19.0
75	SRC	62.0	93.2	57.8	45.8	12.0
76	TACR2	70.3	93.2	65.5	67.0	1.5
77	TBXA2R	75.2	93.2	70.1	67.5	2.6
78	TEK	64.9	93.2	60.5	71.1	10.6

Table A8: A comparison of the predicted values obtained using the similarity calculations with the accuracies of the machine learning models for selected human biological targets. The predicted test accuracy was calculated by multiplying the training accuracy by the similarity between the datasets. A similarity threshold of 0.20 was used for the calculations in the algorithm and a fingerprint length of 10000 bits was used. The similarities were calculated using a sample size of 500 for both the training and test sets. The results were calculated using a neural network architecture of [100,100] and a model trained on the hERG training data. The test accuracies of the targets were obtained by using the hERG model to predict on the various targets in the table.

No.	Target	Calculated similarity between training and test datasets (%)	Model training accuracy (%)	Predicted test accuracy (%)	Model test accuracy (%)	Absolute difference (%)
1	AChE	68.4	93.3	63.8	67.5	3.6
2	ADORA2A	64.9	93.3	60.6	52.6	8.0
3	ADRA2A	73.6	93.3	68.7	77.1	8.4
4	ADRB1	69.9	93.3	65.2	72.3	7.0
5	ADRB2	72.8	93.3	67.9	66.9	1.1
6	AGTR1	61.2	93.3	57.1	65.0	7.8
7	AKT1	68.2	93.3	63.7	71.6	8.0
8	AR	60.7	93.3	56.7	72.8	16.2
9	AVPR1A	73.1	93.3	68.3	81.8	13.6
10	BACE1	59.5	93.3	55.5	63.2	7.7
11	BCHE	68.2	93.3	63.7	72.5	8.8
12	CASP1	66.6	93.3	62.2	63.4	1.2
13	CASP3	66.7	93.3	62.3	61.4	0.8
14	CASP8	55.8	93.3	52.1	75.7	23.6
15	CHRM1	73.7	93.3	68.8	71.5	2.7
16	CHRM2	69.1	93.3	64.5	81.9	17.4
17	CHRM3	73.1	93.3	68.2	77.5	9.3
18	CHRM5	71.4	93.3	66.6	81.4	14.8
19	CHUK	66.9	93.3	62.4	81.7	19.3
20	CSF1R	70.2	93.3	65.5	72.8	7.3
21	CSNK1D	63.1	93.3	58.9	73.2	14.3
22	DRD1	70.4	93.3	65.7	74.9	9.2
23	DRD2	81.0	93.3	75.6	82.8	7.2
24	EDNRA	60.8	93.3	56.7	56.0	0.7
25	EDNRB	60.6	93.3	56.6	65.1	8.5
26	ELANE	66.4	93.3	62.0	52.4	9.6
27	EPHA2	64.8	93.3	60.5	83.3	22.9
28	FGFR1	65.6	93.3	61.2	66.5	5.3
29	FKBP1A	62.3	93.3	58.2	81.8	23.6
30	FLT1	69.7	93.3	65.1	78.9	13.9
31	FLT4	67.1	93.3	62.7	75.0	12.3
32	FYN	61.0	93.3	57.0	83.5	26.5
33	GSK3B	69.8	93.3	65.1	62.1	3.1
34	HDAC3	65.0	93.3	60.7	55.3	5.3
35	HRH1	80.6	93.3	75.2	73.7	1.5
36	HTR2A	80.5	93.3	75.1	76.0	0.9
37	HTR3A	71.6	93.3	66.8	86.9	20.0
38	IGF1R	70.6	93.3	65.9	67.6	1.7
39	INSR	72.6	93.3	67.8	86.0	18.3
40	KDR	67.7	93.3	63.2	58.2	5.0
41	LCK	69.2	93.3	64.6	59.0	5.6
42	LTB4R	58.6	93.3	54.7	72.4	17.6

43	LYN	65.6	93.3	61.2	84.7	23.5
44	MAPK1	72.1	93.3	67.3	52.4	14.8
45	MAPK9	72.0	93.3	67.2	70.4	3.2
46	MAPKAPK2	64.3	93.3	60.0	75.8	15.8
47	MET	69.1	93.3	64.5	63.6	0.9
48	MMP13	70.0	93.3	65.3	44.9	20.4
49	MMP2	68.5	93.3	63.9	44.4	19.6
50	MMP3	64.2	93.3	59.9	43.7	16.2
51	MMP9	66.4	93.3	62.0	48.5	13.5
52	NEK2	65.2	93.3	60.8	84.8	24.0
53	NR3C1	61.6	93.3	57.5	72.8	15.3
54	OPRD1	69.2	93.3	64.6	71.0	6.5
55	OPRM1	75.8	93.3	70.7	75.9	5.2
56	P2RY1	59.3	93.3	55.3	85.4	30.0
57	PAK4	59.1	93.3	55.2	80.1	24.9
58	PDE4A	60.0	93.3	56.0	67.8	11.8
59	PDE5A	70.9	93.3	66.2	70.4	4.2
60	PIK3CA	66.4	93.3	62.0	56.1	5.9
61	PPARG	64.1	93.3	59.8	65.0	5.2
62	PTPN1	59.5	93.3	55.5	62.5	7.0
63	PTPN11	61.2	93.3	57.1	78.7	21.6
64	PTPN2	58.9	93.3	55.0	75.4	20.4
65	RAF1	68.7	93.3	64.1	65.3	1.2
66	RARA	57.1	93.3	53.3	84.3	31.0
67	RARB	53.3	93.3	49.7	87.5	37.7
68	ROCK1	73.1	93.3	68.2	62.0	6.2
69	RPS6KA5	60.8	93.3	56.7	82.5	25.8
70	SERT	78.8	93.3	73.5	79.4	5.8
71	SIRT2	66.0	93.3	61.6	76.5	14.9
72	SIRT3	61.0	93.3	56.9	85.7	28.7
73	SLC6A2	76.2	93.3	71.1	77.2	6.1
74	SLC6A3	74.3	93.3	69.3	77.6	8.3
75	SRC	70.5	93.3	65.8	69.3	3.5
76	TACR2	67.7	93.3	63.2	91.2	28.0
77	TBXA2R	60.7	93.3	56.7	73.1	16.5
78	TEK	61.2	93.3	57.1	73.0	15.9

Table A9: A comparison of the predicted values obtained using the similarity calculations with the accuracies of the machine learning models for selected human biological targets. The predicted test accuracy was calculated by multiplying the training accuracy by the similarity between the datasets. A similarity threshold of 0.20 was used for the calculations in the algorithm and a fingerprint length of 10000 bits was used. The similarities were calculated using a sample size of 500 for both the training and test sets. The results were calculated using

a neural network architecture of [100,100] and a model trained on the SERT training data. The test accuracies of the targets were obtained by using the SERT model to predict on the various targets in the table.

No.	Target	Calculated similarity between training and test datasets (%)	Model training accuracy (%)	Predicted test accuracy (%)	Model test accuracy (%)	Absolute difference (%)
1	AChE	67.8	98.3	66.6	68.1	1.4
2	ADORA2A	56.5	98.3	55.5	46.6	8.9
3	ADRA2A	79.1	98.3	77.8	88.0	10.2
4	ADRB1	71.9	98.3	70.7	85.3	14.6
5	ADRB2	70.0	98.3	68.8	75.5	6.7
6	AGTR1	52.5	98.3	51.6	64.2	12.6
7	AKT1	67.1	98.3	66.0	82.4	16.4
8	AR	53.4	98.3	52.5	70.6	18.2
9	AVPR1A	71.5	98.3	70.2	81.1	10.9
10	BACE1	51.2	98.3	50.3	59.0	8.7
11	BCHE	64.8	98.3	63.7	74.8	11.1
12	CASP1	54.4	98.3	53.5	59.2	5.7
13	CASP3	58.9	98.3	57.9	65.8	7.9
14	CASP8	45.8	98.3	45.0	74.4	29.4
15	CHRM1	80.0	98.3	78.6	73.1	5.5
16	CHRM2	79.6	98.3	78.2	82.8	4.6
17	CHRM3	77.9	98.3	76.6	79.1	2.5
18	CHRM5	83.5	98.3	82.1	84.6	2.5
19	CHUK	56.9	98.3	55.9	80.6	24.6
20	CSF1R	57.2	98.3	56.2	60.0	3.8
21	CSNK1D	59.0	98.3	58.0	65.4	7.5
22	DRD1	77.3	98.3	76.0	84.6	8.6
23	DRD2	88.2	98.3	86.7	94.8	8.1
24	EDNRA	57.5	98.3	56.5	59.6	3.1
25	EDNRB	56.1	98.3	55.1	68.2	13.1
26	ELANE	49.9	98.3	49.0	49.6	0.6
27	EPHA2	53.3	98.3	52.4	83.3	31.0
28	FGFR1	54.6	98.3	53.7	61.1	7.4
29	FKBP1A	60.0	98.3	59.0	87.5	28.5
30	FLT1	59.6	98.3	58.6	71.3	12.7
31	FLT4	57.9	98.3	56.9	68.0	11.0
32	FYN	55.6	98.3	54.6	77.6	22.9
33	GSK3B	66.4	98.3	65.3	48.1	17.2
34	HDAC3	58.4	98.3	57.4	56.1	1.3
35	hERG	77.8	98.3	76.5	67.2	9.2
36	HRH1	83.9	98.3	82.5	87.1	4.6
37	HTR2A	90.8	98.3	89.3	90.8	1.6
38	HTR3A	70.6	98.3	69.4	84.0	14.6

39	IGF1R	56.9	98.3	55.9	75.3	19.4
40	INSR	56.3	98.3	55.3	75.7	20.4
41	KDR	58.8	98.3	57.8	41.1	16.7
42	LCK	60.9	98.3	59.9	45.2	14.6
43	LTB4R	52.9	98.3	52.0	72.6	20.7
44	LYN	56.8	98.3	55.8	73.8	17.9
45	MAPK1	53.4	98.3	52.5	58.3	5.8
46	MAPK9	60.6	98.3	59.6	59.3	0.2
47	MAPKAPK2	59.0	98.3	58.0	76.6	18.6
48	MET	59.7	98.3	58.7	52.4	6.3
49	MMP13	68.3	98.3	67.1	52.8	14.4
50	MMP2	68.9	98.3	67.7	51.4	16.4
51	MMP3	67.3	98.3	66.2	55.2	11.0
52	MMP9	66.0	98.3	64.9	55.8	9.0
53	NEK2	53.6	98.3	52.7	77.9	25.3
54	NR3C1	59.5	98.3	58.5	70.2	11.7
55	OPRD1	72.3	98.3	71.1	86.6	15.5
56	OPRM1	80.6	98.3	79.2	85.9	6.7
57	P2RY1	52.6	98.3	51.7	78.5	26.8
58	PAK4	51.9	98.3	51.0	84.0	33.0
59	PDE4A	57.0	98.3	56.1	67.3	11.3
60	PDE5A	60.3	98.3	59.3	62.7	3.4
61	PIK3CA	46.7	98.3	45.9	47.1	1.2
62	PPARG	62.4	98.3	61.3	60.7	0.6
63	PTPN1	46.4	98.3	45.6	60.9	15.3
64	PTPN11	58.4	98.3	57.4	78.4	21.0
65	PTPN2	50.1	98.3	49.3	78.3	29.0
66	RAF1	53.0	98.3	52.1	55.6	3.5
67	RARA	49.7	98.3	48.9	86.0	37.2
68	RARB	46.5	98.3	45.7	88.8	43.1
69	ROCK1	65.5	98.3	64.4	65.3	0.9
70	RPS6KA5	54.7	98.3	53.8	81.9	28.1
71	SIRT2	56.0	98.3	55.0	74.5	19.5
72	SIRT3	54.7	98.3	53.8	84.7	30.9
73	SLC6A2	91.2	98.3	89.6	93.6	4.0
74	SLC6A3	89.0	98.3	87.5	90.7	3.3
75	SRC	60.8	98.3	59.8	59.3	0.5
76	TACR2	75.3	98.3	74.0	93.6	19.6
77	TBXA2R	58.8	98.3	57.8	74.4	16.6
78	TEK	53.1	98.3	52.2	67.2	15.0

Table A10: Full results for the raw data and curated data for the AChE model vs. various targets. The accuracies were calculated using code taken from the study using a model architecture of [100,100], and by using the model trained on the AChE training data to predict

on the test data of the various targets. A total of 10 runs were performed and the data was shuffled between each run.

	Raw data non-curated data)				Curated data			
Run no.	Test target	Training accuracy (%)	Test accuracy (%)	Average test accuracy (%)	Test target	Training accuracy (%)	Test accuracy (%)	Average test accuracy (%)
1	AChE	94.3	86.3		AChE	94.8	86.6	
2	AChE	94.1	86.1		AChE	94.3	88.9	
3	AChE	94.0	86.4		AChE	95.5	87.4	
4	AChE	94.7	86.6		AChE	95.6	86.9	
5	AChE	95.0	87.4	86.4	AChE	95.2	88.0	87.9
6	AChE	94.9	87.0		AChE	95.8	88.3	
7	AChE	94.7	86.0		AChE	94.7	88.1	
8	AChE	94.0	85.4		AChE	95.0	87.3	
9	AChE	95.2	86.5		AChE	95.4	88.7	
10	AChE	95.3	86.8		AChE	95.1	88.7	
1	ADORA2A	94.3	54.3		ADORA2A	94.1	85.6	
2	ADORA2A	94.1	59.0		ADORA2A	94.9	85.7	
3	ADORA2A	94.0	59.3		ADORA2A	94.9	88.5	
4	ADORA2A	94.7	51.9		ADORA2A	95.1	89.2	
5	ADORA2A	95.0	55.3	56.4	ADORA2A	95.1	88.9	87.8
6	ADORA2A	94.9	54.8		ADORA2A	95.5	87.9	
7	ADORA2A	94.7	59.7		ADORA2A	95.4	89.5	
8	ADORA2A	94.0	60.1		ADORA2A	95.2	87.2	
9	ADORA2A	95.2	52.9		ADORA2A	94.3	86.3	
10	ADORA2A	95.3	56.4		ADORA2A	94.0	89.6	
1	AR	94.3	73.2		AR	95.2	88.3	
2	AR	94.1	72.0		AR	95.0	85.5	
3	AR	94.0	73.1		AR	94.4	87.5	
4	AR	94.7	73.3		AR	95.0	85.3	
5	AR	95.0	74.2	73.0	AR	95.0	88.3	86.7
6	AR	94.9	73.0		AR	95.7	86.5	
7	AR	94.7	72.5		AR	94.9	87.6	
8	AR	94.0	73.2		AR	95.2	85.3	
9	AR	95.2	72.8		AR	94.7	86.1	
10	AR	95.3	73.1		AR	95.0	86.3	
1	hERG	94.3	54.1		hERG	94.7	85.7	
2	hERG	94.1	55.1		hERG	94.8	86.8	
3	hERG	94.0	58.4		hERG	95.6	87.1	
4	hERG	94.7	53.5		hERG	94.8	87.1	
5	hERG	95.0	54.0	55.3	hERG	94.7	88.5	87.2
6	hERG	94.9	51.6		hERG	95.7	86.3	
7	hERG	94.7	58.6		hERG	94.9	88.0	
8	hERG	94.0	59.7		hERG	95.3	85.0	
9	hERG	95.2	53.9		hERG	94.8	87.5	
10	hERG	95.3	54.3		hERG	94.7	90.0	
1	SERT	94.3	51.7	54.3	SERT	95.3	89.0	88.5

2	SERT	94.1	58.7		SERT	95.0	88.8	
3	SERT	94.0	56.7		SERT	95.4	89.8	
4	SERT	94.7	51.6		SERT	95.4	89.7	
5	SERT	95.0	53.0		SERT	95.0	87.4	
6	SERT	94.9	45.9		SERT	95.0	86.2	
7	SERT	94.7	62.4		SERT	95.4	87.0	
8	SERT	94.0	63.2		SERT	95.1	88.7	
9	SERT	95.2	49.1		SERT	94.7	88.4	
10	SERT	95.3	50.4		SERT	95.3	89.8	

Table A11: Number of positive and negative labels in the raw data and curated data for Table S10.

Raw data (non-curated data)							
Training target	Positive labels	Negative labels	Total	**Test target**	Positive labels	Negative labels	Total
AChE	2004	1467	3471	**AChE**	610	496	1106
				ADORA2A	986	501	1487
				AR	648	1815	2463
				hERG	1278	825	2103
				SERT	986	283	1269
Curated data							
Training target	Positive labels	Negative labels	Total	**Test target**	Positive labels	Negative labels	Total
AChE	1971	1390	3361	**AChE**	606	485	1091
AChE	1623	1261	2884	**ADORA2A**	815	490	1305
AChE	1592	1406	2998	**AR**	495	1693	2188
AChE	1802	1320	3122	**hERG**	1105	753	1858
AChE	1682	1215	2897	**SERT**	905	281	1186

Table A12: Average absolute difference (%) of 390 data points vs. various Tanimoto similarity thresholds for Figure 13 in Chapter 3. The average absolute difference is based on 390 data points for that threshold, of which the five targets (AChE, ADORA2A, AR, hERG, and SERT) are tested against the remaining 78 targets.

SIMILARITY THRESHOLD	AVERAGE ABSOLUTE DIFFERENCE (%)
0.1	24.4
0.15	20.9
0.17	16.0
0.18	13.1
0.19	11.8
0.2	12.0
0.21	14.6

0.22	19.1
0.235	25.9
0.24	28.0
0.3	44.6

Table A13: Standard deviations of the absolute difference of 390 data points vs. various Tanimoto similarity thresholds for Figure 13 in Chapter 3. The standard deviation of the average absolute difference is based on 390 data points for that threshold, of which the five targets (AChE, ADORA2A, AR, hERG, and SERT) are tested against the remaining 78 targets.

SIMILARITY THRESHOLD	STANDARD DEVIATION (%)
0.1	15.0
0.15	14.3
0.17	11.9
0.18	10.3
0.19	8.6
0.2	8.5
0.21	10.2
0.22	11.6
0.235	13.1
0.24	13.0
0.3	13.8

Table A14: Results for Figure 15 in Chapter 3, which is a plot of the average absolute difference (%) of 390 data points vs. various sample sizes. The data was shuffled between each run.

Sample size	Run	Test target	Calculated similarity between training and test datasets (%)	Training accuracy (%)	Predicted test accuracy by method (%)	Test accuracy (%)	Absolute difference (%)	Average absolute difference (%)	Standard deviation (%)	Average absolute difference (%) of 3 runs	Standard deviation (%) of 3 runs
10	1	AChE	45.0	94.8	42.6	86.5	43.9	56.9	10.7	56.6	11.2
		ADORA2A	0.0	94.8	0.0	57.4	57.4				
		AR	0.0	94.8	0.0	73.6	73.6				
		hERG	0.0	94.8	0.0	55.2	55.2				
		SERT	0.0	94.8	0.0	54.5	54.5				
	2	AChE	30.0	94.8	28.4	86.5	58.1	51.2	9.6		
		ADORA2A	10.0	94.8	9.5	57.4	48.0				

		AR	10.0	94.8	9.5	73.6	64.1				
		hERG	15.0	94.8	14.2	55.2	41.0				
		SERT	10.0	94.8	9.5	54.5	45.1				
		AChE	10.0	94.8	9.5	86.5	77.1				
		ADORA2A	10.0	94.8	9.5	57.4	48.0				
	3	AR	0.0	94.8	0.0	73.6	73.6	61.7	12.8		
		hERG	0.0	94.8	0.0	55.2	55.2				
		SERT	0.0	94.8	0.0	54.5	54.5				
		AChE	50.0	94.8	47.4	86.5	39.1				
		ADORA2A	28.0	94.8	26.5	57.4	30.9				
	1	AR	25.0	94.8	23.7	73.6	49.9	34.8	10.9		
		hERG	37.0	94.8	35.1	55.2	20.1				
		SERT	22.0	94.8	20.9	54.5	33.7				
		AChE	45.0	94.8	42.6	86.5	43.9				
		ADORA2A	17.0	94.8	16.1	57.4	41.3				
50	2	AR	27.0	94.8	25.6	73.6	48.0	39.1	8.2	37.5	10.7
		hERG	21.0	94.8	19.9	55.2	35.3				
		SERT	29.0	94.8	27.5	54.5	27.0				
		AChE	60.0	94.8	56.9	86.5	29.7				
		ADORA2A	19.0	94.8	18.0	57.4	39.4				
	3	AR	12.0	94.8	11.4	73.6	62.2	38.5	14.2		
		hERG	21.0	94.8	19.9	55.2	35.3				
		SERT	30.0	94.8	28.4	54.5	26.1				
		AChE	70.0	94.8	66.3	86.5	20.2				
		ADORA2A	42.5	94.8	40.3	57.4	17.2				
	1	AR	40.5	94.8	38.4	73.6	35.2	20.3	8.7		
		hERG	45.0	94.8	42.6	55.2	12.6				
		SERT	40.0	94.8	37.9	54.5	16.6				
		AChE	66.0	94.8	62.6	86.5	24.0				
		ADORA2A	46.5	94.8	44.1	57.4	13.4				
100	2	AR	38.5	94.8	36.5	73.6	37.1	21.3	9.8	22.6	9.3
		hERG	39.0	94.8	37.0	55.2	18.2				
		SERT	43.0	94.8	40.8	54.5	13.8				
		AChE	62.0	94.8	58.8	86.5	27.8				
		ADORA2A	34.0	94.8	32.2	57.4	25.2				
	3	AR	33.5	94.8	31.7	73.6	41.9	26.0	10.2		
		hERG	36.0	94.8	34.1	55.2	21.1				
		SERT	42.5	94.8	40.3	54.5	14.3				
		AChE	81.0	94.8	76.8	86.5	9.8				
		ADORA2A	47.3	94.8	44.9	57.4	12.6				
	1	AR	46.3	94.8	43.9	73.6	29.7	12.6	10.4		
150		hERG	48.7	94.8	46.1	55.2	9.1			14.3	9.3
		SERT	55.7	94.8	52.8	54.5	1.8				
	2	AChE	81.7	94.8	77.4	86.5	9.1	11.9	7.2		

		ADORA2A	49.0	94.8	46.4	57.4	11.0				
		AR	52.0	94.8	49.3	73.6	24.3				
		hERG	48.7	94.8	46.1	55.2	9.1				
		SERT	51.3	94.8	48.7	54.5	5.9				
	3	AChE	73.7	94.8	69.8	86.5	16.7	18.3	10.7		
		ADORA2A	43.0	94.8	40.8	57.4	16.7				
		AR	39.3	94.8	37.3	73.6	36.3				
		hERG	43.3	94.8	41.1	55.2	14.1				
		SERT	49.3	94.8	46.8	54.5	7.8				
250	1	AChE	86.0	94.8	81.5	86.5	5.0	7.5	7.1	8.6	7.1
		ADORA2A	57.6	94.8	54.6	57.4	2.8				
		AR	57.2	94.8	54.2	73.6	19.4				
		hERG	60.2	94.8	57.1	55.2	1.8				
		SERT	66.4	94.8	62.9	54.5	8.4				
	2	AChE	84.0	94.8	79.6	86.5	6.9	9.5	8.7		
		ADORA2A	50.6	94.8	48.0	57.4	9.5				
		AR	52.0	94.8	49.3	73.6	24.3				
		hERG	62.2	94.8	59.0	55.2	3.7				
		SERT	60.8	94.8	57.6	54.5	3.1				
	3	AChE	81.0	94.8	76.8	86.5	9.8	8.8	6.8		
		ADORA2A	57.8	94.8	54.8	57.4	2.7				
		AR	56.6	94.8	53.6	73.6	20.0				
		hERG	65.2	94.8	61.8	55.2	6.6				
		SERT	62.6	94.8	59.3	54.5	4.8				
350	1	AChE	86.4	94.8	81.9	86.5	4.6	7.8	6.5	7.6	6.3
		ADORA2A	63.4	94.8	60.1	57.4	2.7				
		AR	57.6	94.8	54.6	73.6	19.0				
		hERG	65.9	94.8	62.4	55.2	7.2				
		SERT	63.4	94.8	60.1	54.5	5.6				
	2	AChE	90.9	94.8	86.1	86.5	0.4	7.9	7.2		
		ADORA2A	63.3	94.8	60.0	57.4	2.5				
		AR	58.3	94.8	55.2	73.6	18.4				
		hERG	69.9	94.8	66.2	55.2	11.0				
		SERT	65.0	94.8	61.6	54.5	7.1				
	3	AChE	88.9	94.8	84.2	86.5	2.3	7.2	6.9		
		ADORA2A	60.0	94.8	56.9	57.4	0.6				
		AR	58.6	94.8	55.5	73.6	18.1				
		hERG	67.3	94.8	63.8	55.2	8.6				
		SERT	64.1	94.8	60.8	54.5	6.3				
500	1	AChE	91.0	94.8	86.2	86.5	0.3	9.1	6.5	8.9	5.5
		ADORA2A	65.3	94.8	61.9	57.4	4.5				
		AR	61.1	94.8	57.9	73.6	15.7				
		hERG	70.9	94.8	67.2	55.2	12.0				
		SERT	71.3	94.8	67.6	54.5	13.0				

	2	AChE	92.6	94.8	87.8	86.5	1.2	8.7	5.3		
		ADORA2A	66.6	94.8	63.1	57.4	5.7				
		AR	65.0	94.8	61.6	73.6	12.0				
		hERG	69.0	94.8	65.4	55.2	10.2				
		SERT	72.9	94.8	69.1	54.5	14.6				
	3	AChE	92.4	94.8	87.6	86.5	1.0	8.8	6.2		
		ADORA2A	65.2	94.8	61.8	57.4	4.4				
		AR	67.7	94.8	64.2	73.6	9.4				
		hERG	72.7	94.8	68.9	55.2	13.7				
		SERT	74.1	94.8	70.2	54.5	15.7				
1000	1	AChE[a]	96.2	94.8	91.1	86.5	4.6	13.2	10.3	13.1	9.1
		ADORA2A[a]	79.1	94.8	74.9	57.4	17.5				
		AR	77.6	94.8	73.5	73.6	0.1				
		hERG	79.9	94.8	75.7	55.2	20.5				
		SERT[a]	82.2	94.8	77.9	54.5	23.3				
	2	AChE[a]	95.4	94.8	90.4	86.5	3.8	13.5	9.9		
		ADORA2A[a]	77.1	94.8	73.0	57.4	15.6				
		AR	75.0	94.8	71.0	73.6	2.6				
		hERG	81.4	94.8	77.1	55.2	21.9				
		SERT[a]	82.4	94.8	78.0	54.5	23.5				
	3	AChE[a]	95.3	94.8	90.3	86.5	3.8	12.7	9.2		
		ADORA2A[a]	75.7	94.8	71.7	57.4	14.3				
		AR	74.8	94.8	70.8	73.6	2.8				
		hERG	78.9	94.8	74.7	55.2	19.5				
		SERT[a]	82.0	94.8	77.7	54.5	23.1				
1500	1	AChE[a]	96.4	94.8	91.3	86.5	4.8	15.5	10.9	15.7	10.2
		ADORA2A[a]	81.3	94.8	77.1	57.4	19.7				
		AR	80.6	94.8	76.4	73.6	2.8				
		hERG	84.7	94.8	80.3	55.2	25.1				
		SERT[a]	84.1	94.8	79.7	54.5	25.2				
	2	AChE[a]	96.6	94.8	91.6	86.5	5.1	16.0	10.7		
		ADORA2A[a]	81.8	94.8	77.5	57.4	20.1				
		AR	81.9	94.8	77.7	73.6	4.0				
		hERG	84.7	94.8	80.2	55.2	25.0				
		SERT[a]	85.0	94.8	80.6	54.5	26.0				
	3	AChE[a]	96.4	94.8	91.4	86.5	4.9	15.7	11.3		
		ADORA2A[a]	82.0	94.8	77.7	57.4	20.3				
		AR	80.0	94.8	75.8	73.6	2.2				
		hERG	84.1	94.8	79.7	55.2	24.5				
		SERT[a]	85.4	94.8	80.9	54.5	26.4				
2000	1	AChE[a]	96.9	94.8	91.8	86.5	5.3	17.4	11.0	17.4	10.1
		ADORA2A[a]	82.5	94.8	78.2	57.4	20.8				
		AR[a]	84.1	94.8	79.7	73.6	6.0				
		hERG[a]	86.8	94.8	82.2	55.2	27.0				

		SERT[a]	86.7	94.8	82.2	54.5	27.7				
	2	AChE[a]	97.0	94.8	91.9	86.5	5.4	17.4	10.9		
		ADORA2A[a]	83.0	94.8	78.7	57.4	21.2				
		AR[a]	84.2	94.8	79.8	73.6	6.2				
		hERG[a]	86.9	94.8	82.3	55.2	27.1				
		SERT[a]	86.2	94.8	81.7	54.5	27.1				
	3	AChE[a]	97.0	94.8	92.0	86.5	5.4	17.4	10.8		
		ADORA2A[a]	83.0	94.8	78.7	57.4	21.2				
		AR[a]	84.3	94.8	79.9	73.6	6.3				
		hERG[a]	87.0	94.8	82.4	55.2	27.2				
		SERT[a]	86.0	94.8	81.5	54.5	27.0				
2500	1	AChE[a]	97.3	94.8	92.2	86.5	5.6	18.3	10.8	18.5	10.0
		ADORA2A[a]	84.2	94.8	79.8	57.4	22.4				
		AR[a]	86.1	94.8	81.6	73.6	7.9				
		hERG[a]	87.8	94.8	83.2	55.2	28.0				
		SERT[a]	86.7	94.8	82.2	54.5	27.6				
	2	AChE[a]	97.6	94.8	92.5	86.5	6.0	18.7	10.8		
		ADORA2A[a]	84.3	94.8	79.9	57.4	22.5				
		AR[a]	86.6	94.8	82.1	73.6	8.5				
		hERG[a]	88.4	94.8	83.8	55.2	28.6				
		SERT[a]	87.3	94.8	82.7	54.5	28.2				
	3	AChE[a]	97.3	94.8	92.2	86.5	5.7	18.4	10.8		
		ADORA2A[a]	83.4	94.8	79.0	57.4	21.6				
		AR[a]	86.5	94.8	82.0	73.6	8.4				
		hERG[a]	87.8	94.8	83.2	55.2	28.0				
		SERT[a]	87.5	94.8	82.9	54.5	28.4				
3000	1	AChE[ab]	97.6	94.8	92.5	86.5	5.9	18.8	10.6	19.0	10.0
		ADORA2A[ab]	84.7	94.8	80.2	57.4	22.8				
		AR[ab]	87.1	94.8	82.5	73.6	8.9				
		hERG[ab]	88.3	94.8	83.7	55.2	28.5				
		SERT[ab]	86.9	94.8	82.4	54.5	27.8				
	2	AChE[ab]	97.6	94.8	92.5	86.5	6.0	19.1	10.9		
		ADORA2A[ab]	84.5	94.8	80.1	57.4	22.6				
		AR[ab]	87.2	94.8	82.6	73.6	9.0				
		hERG[ab]	88.4	94.8	83.8	55.2	28.5				
		SERT[ab]	88.3	94.8	83.7	54.5	29.1				
	3	AChE[ab]	97.7	94.8	92.6	86.5	6.1	19.1	10.7		
		ADORA2A[ab]	84.9	94.8	80.5	57.4	23.0				
		AR[ab]	87.6	94.8	83.0	73.6	9.4				
		hERG[ab]	88.7	94.8	84.1	55.2	28.9				
		SERT[ab]	87.4	94.8	82.8	54.5	28.3				
4000	1	AChE[ab]	97.7	94.8	92.6	86.5	6.1	19.5	11.1	19.5	10.2
		ADORA2A[ab]	85.4	94.8	81.0	57.4	23.5				
		AR[ab]	87.6	94.8	83.0	73.6	9.4				

		hERG[ab]	89.1	94.8	84.5	55.2	29.3				
		SERT[ab]	88.5	94.8	83.8	54.5	29.3				
	2	AChE[ab]	97.7	94.8	92.6	86.5	6.1	19.5	11.1		
		ADORA2A[ab]	85.4	94.8	81.0	57.4	23.5				
		AR[ab]	87.6	94.8	83.0	73.6	9.4				
		hERG[ab]	89.1	94.8	84.5	55.2	29.3				
		SERT[ab]	88.5	94.8	83.8	54.5	29.3				
	3	AChE[ab]	97.7	94.8	92.6	86.5	6.1	19.5	11.1		
		ADORA2A[ab]	85.4	94.8	81.0	57.4	23.5				
		AR[ab]	87.6	94.8	83.0	73.6	9.4				
		hERG[ab]	89.1	94.8	84.5	55.2	29.3				
		SERT[ab]	88.5	94.8	83.8	54.5	29.3				

[a] Test dataset size is smaller than set sample size; all data in test dataset used

[b] Training dataset size is smaller than set sample size; all data in training dataset used

Table A15: A comparison of the predicted values obtained using the similarity calculations with the accuracies of the machine learning models for selected human biological targets. The predicted test accuracy was calculated by multiplying the training accuracy by the similarity between the datasets. A similarity threshold of 0.20 was used for the calculations in the algorithm and a fingerprint length of 10000 bits was used. The similarities were calculated using a sample size of 500 for both the training and test sets. The results were calculated using a neural network architecture of [100,100] and a model trained on the training data for each target (AChE, ADORA2A, AR, hERG, SERT). The test accuracies of the targets were obtained by using the model trained on a target (AChE, ADORA2A, AR, hERG, SERT) to predict on the various targets in the table. The model accuracies were calculated using code taken from the study. Entries 1-79 are the results for AR vs. all other targets, 80-157 (ADORA2A vs. all other targets), 158-235 (AChE vs. all other targets),236-313 (hERG vs. all other targets), 314-391 (SERT vs. all other targets).

No.	Target	Calculated similarity between training and test datasets (%)	Model training accuracy (%)	Predicted test accuracy (%)	Model test accuracy (%)	Absolute difference (%)
1	AChE	47.3	93.2	44.1	49.9	5.8
2	ADORA2A	44.0	93.2	41.0	37.1	3.9
3	ADRA2A	49.9	93.2	46.5	53.3	6.8
4	ADRB1	51.2	93.2	47.7	48.2	0.5
5	ADRB2	47.5	93.2	44.2	50.2	6.0
6	AGTR1	51.1	93.2	47.6	67.5	20.0
7	AKT1	35.0	93.2	32.6	44.5	11.9

9	AVPR1A	49.1	93.2	45.7	66.2	20.5
10	BACE1	37.5	93.2	34.9	38.5	3.6
11	BCHE	57.6	93.2	53.7	61.1	7.4
12	CASP1	43.7	93.2	40.7	66.0	25.3
13	CASP3	49.0	93.2	45.6	59.9	14.2
14	CASP8	55.6	93.2	51.8	75.2	23.4
15	CHRM1	41.6	93.2	38.8	37.2	1.6
16	CHRM2	53.6	93.2	50.0	55.1	5.2
17	CHRM3	45.8	93.2	42.6	42.8	0.1
18	CHRM5	53.5	93.2	49.8	62.7	12.8
19	CHUK	54.7	93.2	51.0	74.4	23.4
20	CSF1R	41.7	93.2	38.9	49.7	10.8
21	CSNK1D	50.1	93.2	46.6	62.0	15.4
22	DRD1	59.9	93.2	55.8	61.2	5.4
23	DRD2	33.0	93.2	30.8	17.3	13.4
24	EDNRA	41.8	93.2	39.0	49.4	10.4
25	EDNRB	45.2	93.2	42.2	58.8	16.7
26	ELANE	44.6	93.2	41.5	61.8	20.3
27	EPHA2	62.8	93.2	58.5	64.7	6.2
28	FGFR1	37.8	93.2	35.2	47.2	12.0
29	FKBP1A	53.8	93.2	50.1	69.0	18.9
30	FLT1	57.2	93.2	53.3	66.3	13.0
31	FLT4	49.3	93.2	46.0	61.4	15.4
32	FYN	54.9	93.2	51.2	75.1	23.9
33	GSK3B	39.5	93.2	36.8	41.7	4.9
34	HDAC3	48.4	93.2	45.1	52.8	7.7
35	hERG	38.8	93.2	36.2	43.3	7.1
36	HRH1	48.8	93.2	45.5	48.2	2.8
37	HTR2A	35.4	93.2	33.0	24.8	8.2
38	HTR3A	53.8	93.2	50.1	68.8	18.6
39	IGF1R	35.7	93.2	33.2	50.4	17.1
40	INSR	43.9	93.2	40.9	66.4	25.5
41	KDR	39.4	93.2	36.7	30.8	5.9
42	LCK	22.4	93.2	20.8	35.0	14.2
43	LTB4R	56.5	93.2	52.6	68.5	15.9
44	LYN	53.0	93.2	49.4	73.5	24.1
45	MAPK1	51.7	93.2	48.2	60.7	12.5
46	MAPK9	44.0	93.2	41.0	53.6	12.6
47	MAPKAPK2	45.9	93.2	42.8	66.2	23.4
48	MET	36.7	93.2	34.2	42.6	8.4
49	MMP13	39.5	93.2	36.8	36.4	0.4
50	MMP2	43.0	93.2	40.1	37.7	2.4
51	MMP3	40.8	93.2	38.0	37.6	0.3
52	MMP9	52.6	93.2	49.0	43.7	5.3

53	NEK2	57.1	93.2	53.2	79.9	26.7
54	NR3C1	83.2	93.2	77.6	83.3	5.8
55	OPRD1	38.2	93.2	35.6	30.4	5.2
56	OPRM1	45.5	93.2	42.4	39.0	3.4
57	P2RY1	52.5	93.2	48.9	66.4	17.5
58	PAK4	55.9	93.2	52.0	76.7	24.6
59	PDE4A	46.7	93.2	43.5	58.5	15.0
60	PDE5A	42.5	93.2	39.6	47.1	7.5
61	PIK3CA	41.6	93.2	38.8	46.7	8.0
62	PPARG	72.6	93.2	67.7	63.6	4.1
63	PTPN1	47.1	93.2	43.9	63.2	19.3
64	PTPN11	55.9	93.2	52.1	76.1	24.0
65	PTPN2	56.2	93.2	52.4	77.5	25.1
66	RAF1	53.1	93.2	49.5	72.5	23.0
67	RARA	73.0	93.2	68.0	86.9	18.9
68	RARB	74.2	93.2	69.1	89.6	20.5
69	ROCK1	43.9	93.2	40.9	52.9	12.0
70	RPS6KA5	58.4	93.2	54.4	80.0	25.6
71	SERT	36.2	93.2	33.7	23.5	10.2
72	SIRT2	54.8	93.2	51.0	74.8	23.7
73	SIRT3	58.3	93.2	54.3	82.1	27.8
74	SLC6A2	48.2	93.2	44.9	42.2	2.7
75	SLC6A3	49.3	93.2	45.9	42.9	3.1
76	SRC	38.5	93.2	35.9	45.8	9.9
77	TACR2	56.1	93.2	52.3	67.0	14.7
78	TBXA2R	59.1	93.2	55.1	67.5	12.4
79	TEK	49.1	93.2	45.7	71.1	25.3
80	AChE	31.1	97.4	30.4	55.9	25.5
81	ADRA2A	43.6	97.4	42.5	62.6	20.1
82	ADRB1	34.9	97.4	34.0	58.4	24.4
83	ADRB2	34.7	97.4	33.8	58.3	24.5
84	AGTR1	41.6	97.4	40.5	86.2	45.7
85	AKT1	30.7	97.4	29.9	82.4	52.5
86	AR	50.0	97.4	48.7	73.4	24.6
87	AVPR1A	41.1	97.4	40.1	65.7	25.7
88	BACE1	26.1	97.4	25.4	50.3	24.9
89	BCHE	42.1	97.4	41.0	71.6	30.6
90	CASP1	35.9	97.4	35.0	57.3	22.4
91	CASP3	39.0	97.4	38.0	58.2	20.2
92	CASP8	46.2	97.4	45.0	76.3	31.3
93	CHRM1	30.1	97.4	29.4	41.9	12.5
94	CHRM2	40.4	97.4	39.4	56.1	16.7
95	CHRM3	33.9	97.4	33.0	51.1	18.1
96	CHRM5	41.6	97.4	40.6	63.3	22.8

97	CHUK	49.3	97.4	48.1	90.4	42.3
98	CSF1R	36.2	97.4	35.3	84.8	49.5
99	CSNK1D	47.1	97.4	45.9	91.5	45.6
100	DRD1	49.5	97.4	48.2	63.9	15.7
101	DRD2	30.8	97.4	30.0	35.6	5.6
102	EDNRA	34.2	97.4	33.4	59.3	25.9
103	EDNRB	35.8	97.4	34.9	68.0	33.2
104	ELANE	30.2	97.4	29.4	46.9	17.6
105	EPHA2	43.8	97.4	42.6	90.8	48.2
106	FGFR1	33.1	97.4	32.3	88.3	56.1
107	FKBP1A	43.6	97.4	42.4	70.1	27.7
108	FLT1	48.9	97.4	47.7	88.0	40.3
109	FLT4	43.7	97.4	42.6	84.3	41.7
110	FYN	48.3	97.4	47.0	90.0	43.0
111	GSK3B	41.6	97.4	40.5	72.6	32.0
112	HDAC3	37.1	97.4	36.1	67.7	31.6
113	hERG	28.1	97.4	27.4	49.2	21.8
114	HRH1	36.6	97.4	35.7	53.1	17.4
115	HTR2A	30.9	97.4	30.1	40.4	10.3
116	HTR3A	46.5	97.4	45.3	81.9	36.6
117	IGF1R	32.7	97.4	31.9	91.3	59.4
118	INSR	38.5	97.4	37.5	87.7	50.1
119	KDR	30.7	97.4	29.9	74.4	44.4
120	LCK	23.1	97.4	22.5	80.3	57.8
121	LTB4R	46.7	97.4	45.5	73.5	28.0
122	LYN	47.0	97.4	45.8	92.6	46.8
123	MAPK1	45.6	97.4	44.4	62.3	17.9
124	MAPK9	46.3	97.4	45.1	90.8	45.7
125	MAPKAPK2	41.8	97.4	40.7	83.4	42.7
126	MET	29.3	97.4	28.6	78.1	49.6
127	MMP13	34.3	97.4	33.5	44.3	10.9
128	MMP2	28.7	97.4	28.0	35.1	7.1
129	MMP3	30.6	97.4	29.8	40.9	11.1
130	MMP9	38.0	97.4	37.1	43.9	6.8
131	NEK2	51.8	97.4	50.5	91.1	40.6
132	NR3C1	47.9	97.4	46.7	72.6	25.9
133	OPRD1	26.8	97.4	26.1	38.8	12.7
134	OPRM1	33.4	97.4	32.5	44.9	12.4
135	P2RY1	47.6	97.4	46.4	81.5	35.2
136	PAK4	47.3	97.4	46.1	92.1	46.0
137	PDE4A	41.3	97.4	40.3	76.7	36.4
138	PDE5A	46.2	97.4	45.0	81.0	36.0
139	PIK3CA	38.0	97.4	37.1	82.5	45.4
140	PPARG	46.4	97.4	45.2	65.0	19.8

141	PTPN1	33.1	97.4	32.3	64.5	32.2
142	PTPN11	47.7	97.4	46.4	77.1	30.7
143	PTPN2	45.1	97.4	43.9	77.0	33.0
144	RAF1	38.1	97.4	37.2	77.3	40.1
145	RARA	55.0	97.4	53.6	87.1	33.6
146	RARB	56.9	97.4	55.5	90.3	34.8
147	ROCK1	40.9	97.4	39.9	80.2	40.3
148	RPS6KA5	51.2	97.4	49.9	92.2	42.3
149	SERT	27.0	97.4	26.3	31.5	5.2
150	SIRT2	43.4	97.4	42.3	77.3	34.9
151	SIRT3	48.8	97.4	47.5	89.6	42.0
152	SLC6A2	34.6	97.4	33.7	44.0	10.4
153	SLC6A3	35.6	97.4	34.7	44.0	9.3
154	SRC	38.6	97.4	37.6	80.8	43.2
155	TACR2	48.2	97.4	47.0	75.6	28.6
156	TBXA2R	45.9	97.4	44.7	69.4	24.7
157	TEK	45.5	97.4	44.3	82.8	38.4
158	ADORA2A	30.7	94.8	29.1	57.4	28.3
159	ADRA2A	42.9	94.8	40.7	73.2	32.5
160	ADRB1	34.6	94.8	32.8	61.2	28.4
161	ADRB2	33.3	94.8	31.5	62.1	30.6
162	AGTR1	38.2	94.8	36.2	65.0	28.8
163	AKT1	26.8	94.8	25.4	52.1	26.7
164	AR	45.9	94.8	43.5	73.6	30.1
165	AVPR1A	39.0	94.8	36.9	71.7	34.8
166	BACE1	32.8	94.8	31.1	46.6	15.6
167	BCHE	77.4	94.8	73.4	80.9	7.5
168	CASP1	32.1	94.8	30.4	58.2	27.8
169	CASP3	36.7	94.8	34.8	64.7	29.9
170	CASP8	41.8	94.8	39.6	70.3	30.7
171	CHRM1	36.6	94.8	34.7	59.3	24.7
172	CHRM2	43.3	94.8	41.0	71.2	30.2
173	CHRM3	40.3	94.8	38.2	66.7	28.5
174	CHRM5	42.7	94.8	40.5	72.4	31.9
175	CHUK	42.7	94.8	40.5	80.0	39.5
176	CSF1R	31.2	94.8	29.6	60.8	31.2
177	CSNK1D	37.1	94.8	35.1	65.7	30.5
178	DRD1	46.2	94.8	43.7	76.0	32.3
179	DRD2	47.1	94.8	44.7	66.1	21.4
180	EDNRA	31.5	94.8	29.8	53.2	23.4
181	EDNRB	33.5	94.8	31.8	58.6	26.9
182	ELANE	28.6	94.8	27.1	50.4	23.3
183	EPHA2	40.5	94.8	38.4	72.7	34.3
184	FGFR1	27.2	94.8	25.8	60.2	34.5

185	FKBP1A	43.3	94.8	41.0	85.1	44.0
186	FLT1	37.9	94.8	36.0	73.3	37.4
187	FLT4	37.9	94.8	35.9	70.9	35.0
188	FYN	42.3	94.8	40.1	78.1	38.0
189	GSK3B	28.9	94.8	27.3	52.9	25.5
190	HDAC3	34.5	94.8	32.7	55.7	23.0
191	hERG	35.2	94.8	33.4	55.2	21.9
192	HRH1	48.0	94.8	45.5	69.2	23.7
193	HTR2A	40.5	94.8	38.4	63.4	25.0
194	HTR3A	48.6	94.8	46.1	84.8	38.7
195	IGF1R	27.2	94.8	25.8	53.9	28.1
196	INSR	33.8	94.8	32.1	66.4	34.3
197	KDR	22.4	94.8	21.2	43.8	22.6
198	LCK	16.2	94.8	15.4	43.0	27.6
199	LTB4R	45.1	94.8	42.7	71.5	28.8
200	LYN	40.2	94.8	38.1	74.3	36.3
201	MAPK1	37.9	94.8	35.9	56.7	20.8
202	MAPK9	35.8	94.8	33.9	68.9	34.9
203	MAPKAPK2	35.5	94.8	33.7	70.6	36.9
204	MET	27.3	94.8	25.9	54.0	28.1
205	MMP13	33.4	94.8	31.6	45.7	14.1
206	MMP2	30.9	94.8	29.3	39.8	10.5
207	MMP3	30.5	94.8	28.9	43.8	14.9
208	MMP9	37.6	94.8	35.7	46.1	10.4
209	NEK2	46.7	94.8	44.3	81.7	37.4
210	NR3C1	45.3	94.8	43.0	71.1	28.2
211	OPRD1	35.4	94.8	33.6	66.2	32.7
212	OPRM1	40.5	94.8	38.4	65.4	27.0
213	P2RY1	42.9	94.8	40.7	72.6	31.9
214	PAK4	43.7	94.8	41.4	82.0	40.6
215	PDE4A	37.4	94.8	35.5	67.6	32.1
216	PDE5A	32.5	94.8	30.8	63.2	32.4
217	PIK3CA	25.7	94.8	24.3	42.1	17.7
218	PPARG	45.6	94.8	43.3	60.9	17.6
219	PTPN1	34.0	94.8	32.2	62.2	30.0
220	PTPN11	42.4	94.8	40.2	73.2	33.1
221	PTPN2	43.3	94.8	41.1	75.4	34.3
222	RAF1	31.2	94.8	29.6	55.9	26.3
223	RARA	47.8	94.8	45.3	83.9	38.5
224	RARB	49.8	94.8	47.2	87.6	40.4
225	ROCK1	33.6	94.8	31.9	61.4	29.5
226	RPS6KA5	47.9	94.8	45.4	82.5	37.1
227	SERT	32.8	94.8	31.1	54.5	23.5
228	SIRT2	41.6	94.8	39.4	80.5	41.1

229	SIRT3	48.2	94.8	45.7	86.6	41.0
230	SLC6A2	35.3	94.8	33.5	54.4	20.9
231	SLC6A3	36.2	94.8	34.4	59.6	25.2
232	SRC	28.0	94.8	26.5	57.5	30.9
233	TACR2	45.6	94.8	43.2	75.2	32.0
234	TBXA2R	38.2	94.8	36.2	69.3	33.1
235	TEK	34.6	94.8	32.8	70.6	37.9
236	AChE	36.5	93.3	34.1	67.5	33.4
237	ADORA2A	34.4	93.3	32.1	52.6	20.5
238	ADRA2A	42.9	93.3	40.1	77.1	37.0
239	ADRB1	46.8	93.3	43.7	72.3	28.6
240	ADRB2	39.3	93.3	36.7	66.9	30.2
241	AGTR1	35.8	93.3	33.4	65.0	31.5
242	AKT1	31.0	93.3	29.0	71.6	42.7
243	AR	42.7	93.3	39.9	72.8	33.0
244	AVPR1A	42.1	93.3	39.3	81.8	42.6
245	BACE1	33.8	93.3	31.5	63.2	31.7
246	BCHE	41.1	93.3	38.4	72.5	34.1
247	CASP1	27.4	93.3	25.6	63.4	37.8
248	CASP3	37.9	93.3	35.3	61.4	26.1
249	CASP8	40.8	93.3	38.0	75.7	37.7
250	CHRM1	46.1	93.3	43.0	71.5	28.5
251	CHRM2	45.9	93.3	42.8	81.9	39.1
252	CHRM3	44.7	93.3	41.7	77.5	35.8
253	CHRM5	47.7	93.3	44.5	81.4	37.0
254	CHUK	47.1	93.3	44.0	81.7	37.7
255	CSF1R	39.6	93.3	36.9	72.8	35.9
256	CSNK1D	42.2	93.3	39.4	73.2	33.8
257	DRD1	49.6	93.3	46.3	74.9	28.7
258	DRD2	62.1	93.3	58.0	82.8	24.8
259	EDNRA	30.9	93.3	28.8	56.0	27.2
260	EDNRB	31.3	93.3	29.2	65.1	35.9
261	ELANE	27.0	93.3	25.2	52.4	27.2
262	EPHA2	46.3	93.3	43.2	83.3	40.1
263	FGFR1	32.8	93.3	30.6	66.5	35.9
264	FKBP1A	44.2	93.3	41.2	81.8	40.6
265	FLT1	44.0	93.3	41.1	78.9	37.8
266	FLT4	42.7	93.3	39.8	75.0	35.2
267	FYN	48.3	93.3	45.1	83.5	38.4
268	GSK3B	34.5	93.3	32.2	62.1	29.9
269	HDAC3	37.3	93.3	34.8	55.3	20.5
270	HRH1	56.5	93.3	52.7	73.7	21.0
271	HTR2A	54.9	93.3	51.2	76.0	24.8
272	HTR3A	53.7	93.3	50.1	86.9	36.8

273	IGF1R	38.1	93.3	35.5	67.6	32.0
274	INSR	46.1	93.3	43.0	86.0	43.0
275	KDR	35.0	93.3	32.6	58.2	25.6
276	LCK	30.5	93.3	28.5	59.0	30.5
277	LTB4R	43.2	93.3	40.3	72.4	32.1
278	LYN	47.9	93.3	44.7	84.7	40.0
279	MAPK1	36.6	93.3	34.2	52.4	18.3
280	MAPK9	39.5	93.3	36.8	70.4	33.6
281	MAPKAPK2	37.2	93.3	34.7	75.8	41.1
282	MET	37.6	93.3	35.1	63.6	28.5
283	MMP13	33.9	93.3	31.6	44.9	13.3
284	MMP2	31.7	93.3	29.6	44.4	14.7
285	MMP3	29.3	93.3	27.4	43.7	16.3
286	MMP9	36.2	93.3	33.8	48.5	14.7
287	NEK2	49.6	93.3	46.3	84.8	38.5
288	NR3C1	42.2	93.3	39.4	72.8	33.5
289	OPRD1	46.3	93.3	43.2	71.0	27.9
290	OPRM1	48.6	93.3	45.4	75.9	30.5
291	P2RY1	44.9	93.3	41.9	85.4	43.4
292	PAK4	46.6	93.3	43.4	80.1	36.6
293	PDE4A	35.9	93.3	33.5	67.8	34.3
294	PDE5A	38.2	93.3	35.7	70.4	34.7
295	PIK3CA	30.6	93.3	28.5	56.1	27.5
296	PPARG	43.7	93.3	40.7	65.0	24.2
297	PTPN1	29.7	93.3	27.8	62.5	34.8
298	PTPN11	45.0	93.3	42.0	78.7	36.7
299	PTPN2	40.4	93.3	37.7	75.4	37.7
300	RAF1	34.4	93.3	32.1	65.3	33.2
301	RARA	45.0	93.3	42.0	84.3	42.3
302	RARB	47.7	93.3	44.5	87.5	43.0
303	ROCK1	37.5	93.3	35.0	62.0	27.0
304	RPS6KA5	47.5	93.3	44.3	82.5	38.2
305	SERT	53.0	93.3	49.5	79.4	29.9
306	SIRT2	43.7	93.3	40.8	76.5	35.7
307	SIRT3	49.2	93.3	45.9	85.7	39.7
308	SLC6A2	50.0	93.3	46.7	77.2	30.6
309	SLC6A3	50.6	93.3	47.2	77.6	30.4
310	SRC	38.1	93.3	35.6	69.3	33.7
311	TACR2	47.3	93.3	44.1	91.2	47.1
312	TBXA2R	39.1	93.3	36.5	73.1	36.7
313	TEK	40.0	93.3	37.4	73.0	35.6
314	AChE	32.3	98.3	31.8	68.1	36.3
315	ADORA2A	24.8	98.3	24.3	46.6	22.3
316	ADRA2A	50.0	98.3	49.2	88.0	38.8

317	ADRB1	41.7	98.3	41.0	85.3	44.3
318	ADRB2	32.5	98.3	31.9	75.5	43.6
319	AGTR1	33.9	98.3	33.3	64.2	30.9
320	AKT1	25.8	98.3	25.4	82.4	57.0
321	AR	40.3	98.3	39.6	70.6	31.0
322	AVPR1A	40.8	98.3	40.1	81.1	41.1
323	BACE1	20.1	98.3	19.8	59.0	39.2
324	BCHE	38.5	98.3	37.9	74.8	37.0
325	CASP1	22.3	98.3	22.0	59.2	37.2
326	CASP3	30.5	98.3	29.9	65.8	35.9
327	CASP8	41.1	98.3	40.4	74.4	34.0
328	CHRM1	40.5	98.3	39.8	73.1	33.3
329	CHRM2	42.8	98.3	42.1	82.8	40.7
330	CHRM3	38.0	98.3	37.3	79.1	41.8
331	CHRM5	50.4	98.3	49.5	84.6	35.1
332	CHUK	43.4	98.3	42.6	80.6	37.9
333	CSF1R	29.5	98.3	29.0	60.0	31.0
334	CSNK1D	36.7	98.3	36.1	65.4	29.3
335	DRD1	51.8	98.3	50.9	84.6	33.7
336	DRD2	59.5	98.3	58.4	94.8	36.4
337	EDNRA	29.4	98.3	28.9	59.6	30.7
338	EDNRB	30.8	98.3	30.3	68.2	38.0
339	ELANE	23.6	98.3	23.2	49.6	26.4
340	EPHA2	38.0	98.3	37.3	83.3	46.0
341	FGFR1	25.3	98.3	24.9	61.1	36.2
342	FKBP1A	41.3	98.3	40.6	87.5	46.9
343	FLT1	36.4	98.3	35.7	71.3	35.5
344	FLT4	36.1	98.3	35.4	68.0	32.5
345	FYN	41.5	98.3	40.8	77.6	36.7
346	GSK3B	28.0	98.3	27.5	48.1	20.6
347	HDAC3	31.4	98.3	30.9	56.1	25.3
348	hERG	50.7	98.3	49.8	67.2	17.4
349	HRH1	56.4	98.3	55.4	87.1	31.7
350	HTR2A	62.0	98.3	60.9	90.8	29.9
351	HTR3A	51.1	98.3	50.2	84.0	33.8
352	IGF1R	25.5	98.3	25.1	75.3	50.2
353	INSR	32.3	98.3	31.8	75.7	43.9
354	KDR	21.1	98.3	20.8	41.1	20.4
355	LCK	16.0	98.3	15.7	45.2	29.5
356	LTB4R	42.5	98.3	41.7	72.6	30.9
357	LYN	39.5	98.3	38.8	73.8	35.0
358	MAPK1	26.6	98.3	26.1	58.3	32.1
359	MAPK9	30.5	98.3	30.0	59.3	29.3
360	MAPKAPK2	34.0	98.3	33.4	76.6	43.2

361	MET	23.6	98.3	23.2	52.4	29.2
362	MMP13	29.1	98.3	28.6	52.8	24.1
363	MMP2	24.6	98.3	24.1	51.4	27.2
364	MMP3	24.8	98.3	24.4	55.2	30.8
365	MMP9	31.9	98.3	31.3	55.8	24.5
366	NEK2	38.7	98.3	38.0	77.9	39.9
367	NR3C1	35.6	98.3	34.9	70.2	35.2
368	OPRD1	40.8	98.3	40.1	86.6	46.5
369	OPRM1	40.0	98.3	39.3	85.9	46.6
370	P2RY1	42.1	98.3	41.4	78.5	37.1
371	PAK4	34.6	98.3	34.0	84.0	49.9
372	PDE4A	28.7	98.3	28.2	67.3	39.1
373	PDE5A	21.7	98.3	21.4	62.7	41.3
374	PIK3CA	36.3	98.3	35.7	47.1	11.4
375	PPARG	25.8	98.3	25.4	60.7	35.3
376	PTPN1	41.8	98.3	41.1	60.9	19.9
377	PTPN11	38.7	98.3	38.1	78.4	40.4
378	PTPN2	27.8	98.3	27.3	78.3	50.9
379	RAF1	45.0	98.3	44.2	55.6	11.4
380	RARA	46.5	98.3	45.7	86.0	40.4
381	RARB	30.7	98.3	30.2	88.8	58.6
382	ROCK1	46.0	98.3	45.2	65.3	20.1
383	RPS6KA5	39.3	98.3	38.7	81.9	43.2
384	SIRT2	46.3	98.3	45.5	74.5	29.0
385	SIRT3	84.7	98.3	83.3	84.7	1.4
386	SLC6A2	82.1	98.3	80.7	93.6	12.9
387	SLC6A3	25.1	98.3	24.6	90.7	66.1
388	SRC	48.0	98.3	47.1	59.3	12.1
389	TACR2	36.1	98.3	35.5	93.6	58.1
390	TBXA2R	33.2	98.3	32.6	74.4	41.8
391	TEK	19.1	98.3	18.8	67.2	48.4

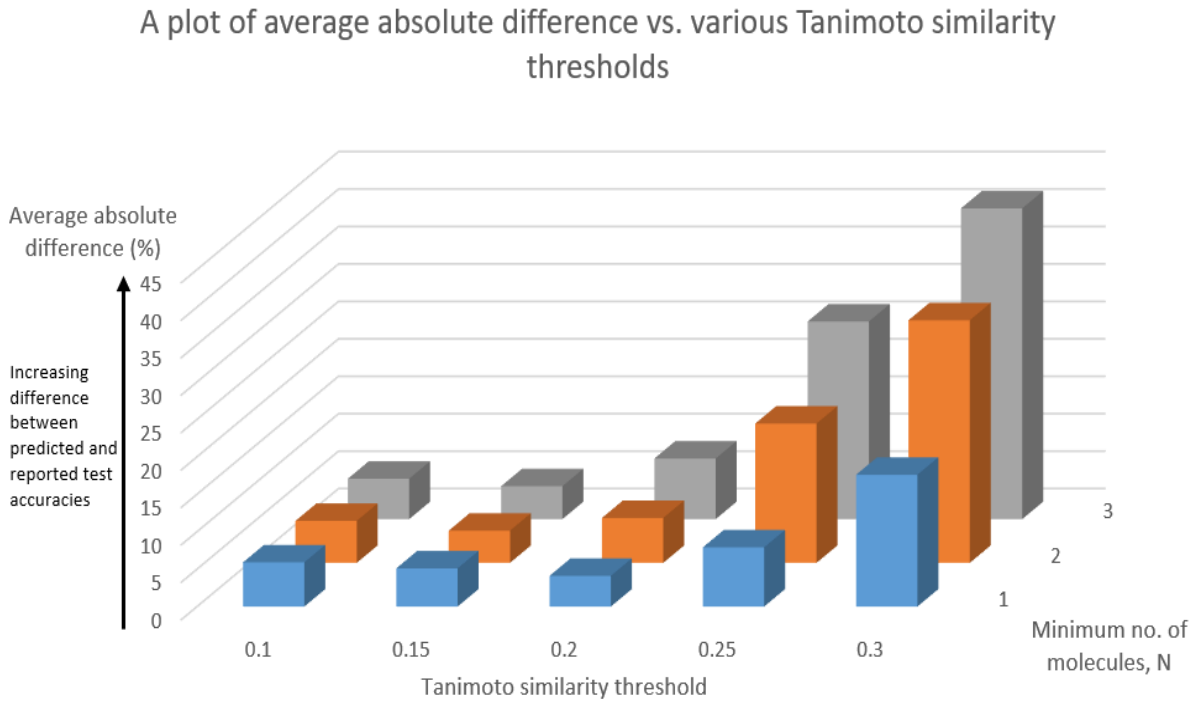


Figure A5: Plot showing the full results for the initial scan of the effect of changing the Tanimoto similarity threshold on the reliability of the transferability process of a ML model to a new dataset. A plot of average absolute difference of five targets vs. the Tanimoto similarity threshold for various Tanimoto similarity thresholds tested during the optimisation process. The average absolute difference refers to the average (of 5 targets) absolute difference between the predicted and model test accuracies of each target (AChE, ADORA2A, AR, hERG, and SERT). An increasing value of the absolute difference means that the method is increasingly less representative of the ML model (method predicts a different outcome as compared to the ML model). The predicted model test accuracies were calculated by multiplying the model training accuracies by the average similarity between datasets for the same target. The reported test accuracies were taken from the paper. All training and test sets use a sample size of 500.

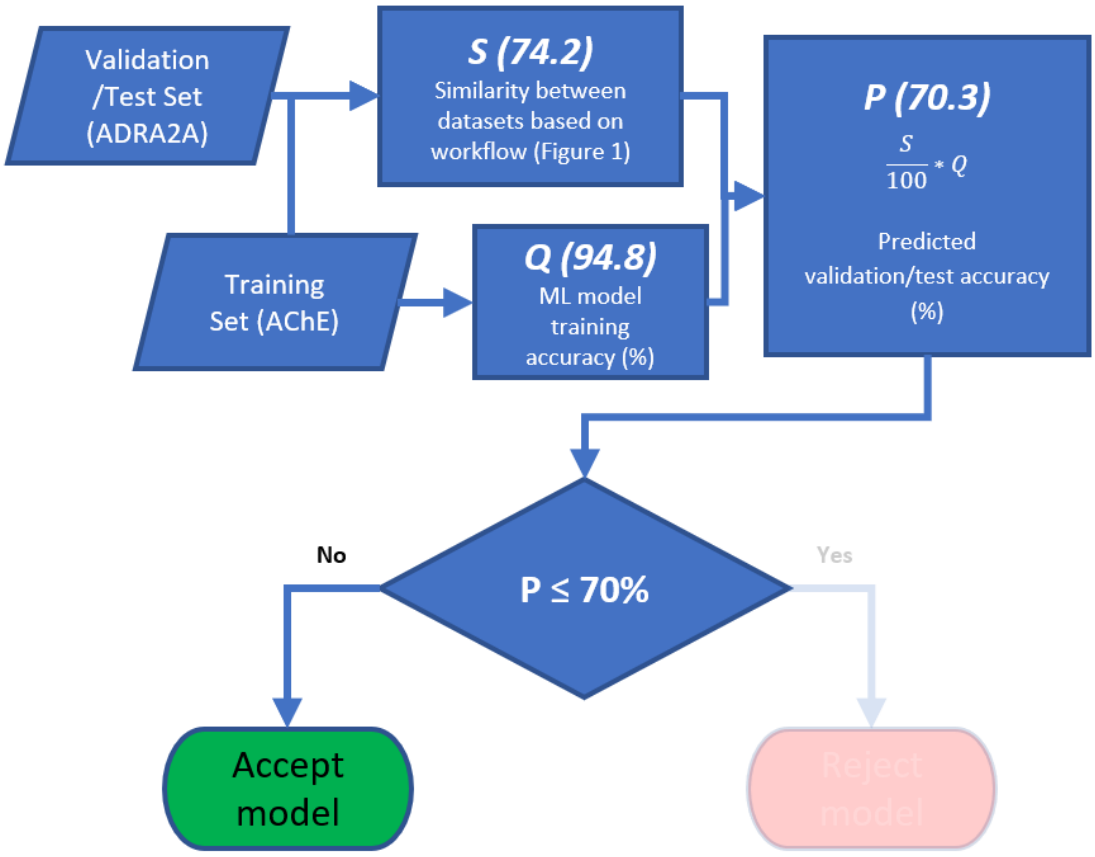


Figure A6: Workflow showing the outcome where the model of dataset 1 can be applied on dataset 2. Data from Table A5 was used to illustrate the outcome. Baseline accuracy refers to the minimum accuracy when an unoptimised model is used *i.e.* the accuracy obtained when it is assumed that for all data points, the model predicts their class/label to be the class/label with the most observations in the dataset.

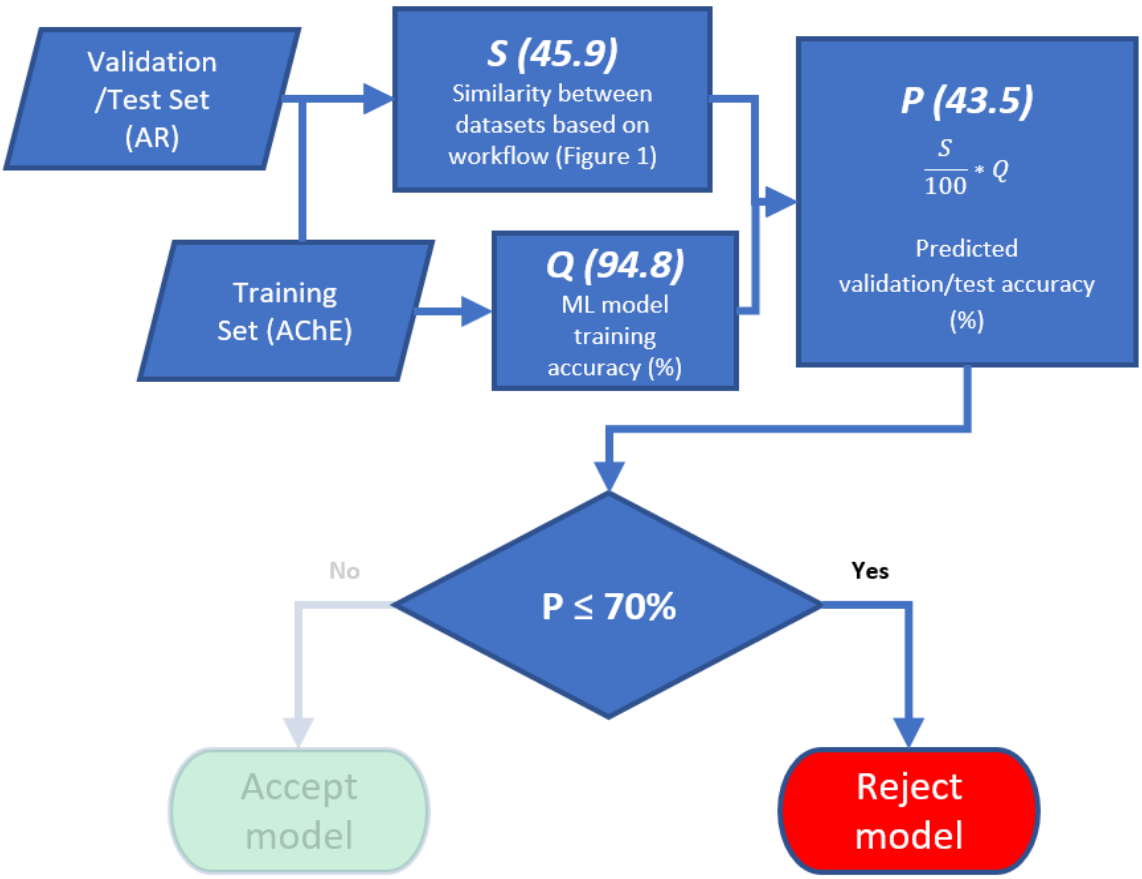


Figure A7: Workflow showing the outcome where the model of dataset 1 cannot be applied on dataset 2. Data from Table A5 was used to illustrate the outcome. Baseline accuracy refers to the minimum accuracy when an unoptimised model is used *i.e.* the accuracy obtained when it is assumed that for all data points, the model predicts their class/label to be the class/label with the most observations in the dataset.

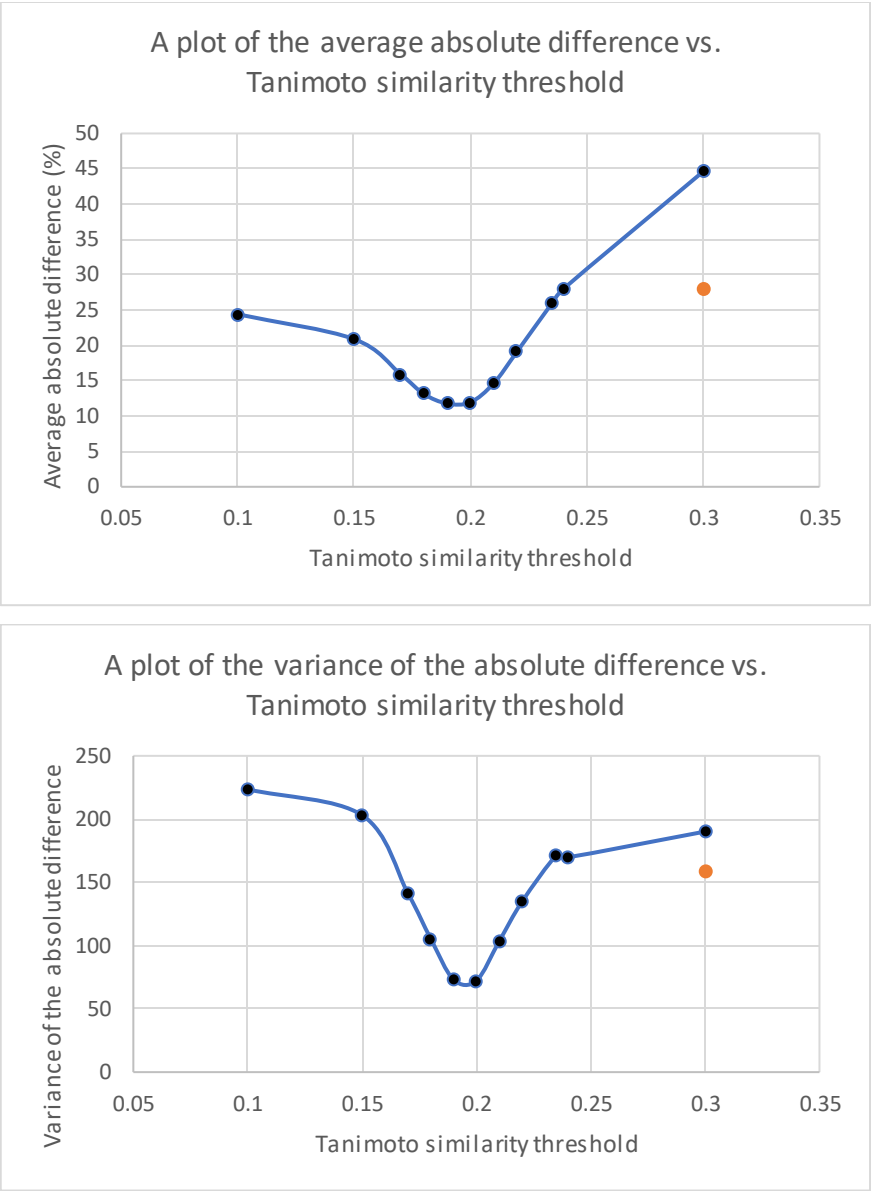


Figure A8: Plots showing the average absolute difference (top) and the variance of the absolute difference (bottom) at various thresholds. The blue points were obtained using a sample size of 500 while the orange point was obtained using all the data. The data for five targets (AChE, ADORA2A, AR, hERG, SERT) was used and is summarised in Table A14, where each of the five targets was tested against all other targets. The two plots in this Figure show that using all data still results in a lower performance with this method as compared to using a smaller sample size and a lower threshold. It is expected that using all data at a threshold of 0.2 will still be worse than if a sample size of 500 is used. This is because the similarity between datasets will be calculated to be higher as there is a higher chance for a similar molecule to be found in the other dataset that is tested. Hence, this would result in a higher average absolute difference and variance.

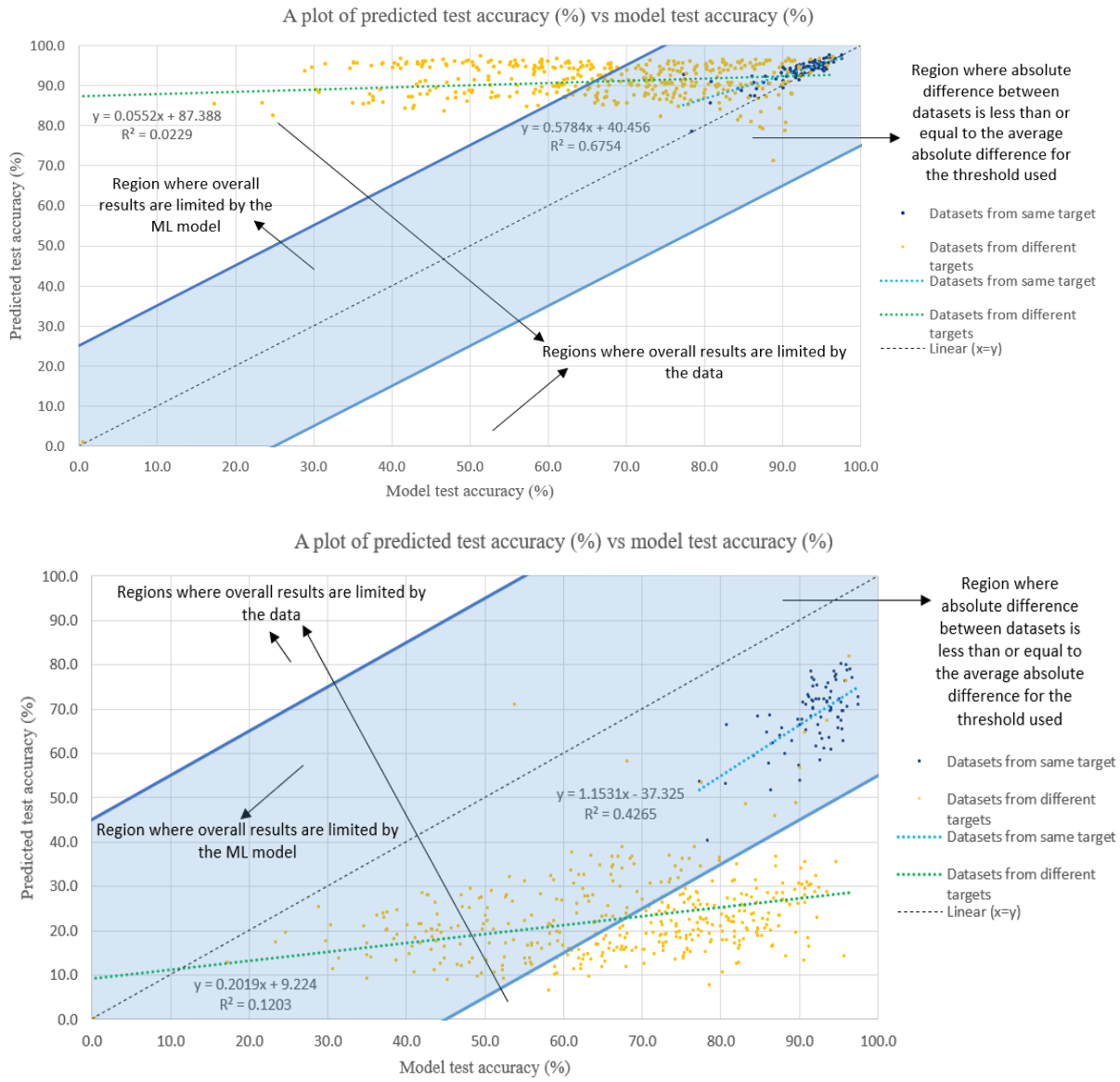


Figure A9: A plot of predicted test accuracy (%) vs. model test accuracy (%) for datasets from various targets. Each point on the plot represents a different combination of training and test datasets from the various targets. When the datasets are from the same target, the training and test sets used come from the same target. When the datasets are from the different targets, the training and test sets used come from different targets. The x = y line has also been plotted and is shown as a black, dashed line in the figure. (Top) The blue region indicates the region where the absolute difference between datasets is less than or equal to the average absolute difference for the threshold (25%) as determined in Figure 3. These settings were used for the calculations: a similarity threshold of 0.10 and a sample size of 500. (Bottom) The blue region indicates the region where the absolute difference between datasets is less than or equal to the average absolute difference for the threshold (45%) as determined in Figure 13. These settings were used for the calculations: a similarity threshold of 0.30 and a sample size of 500.

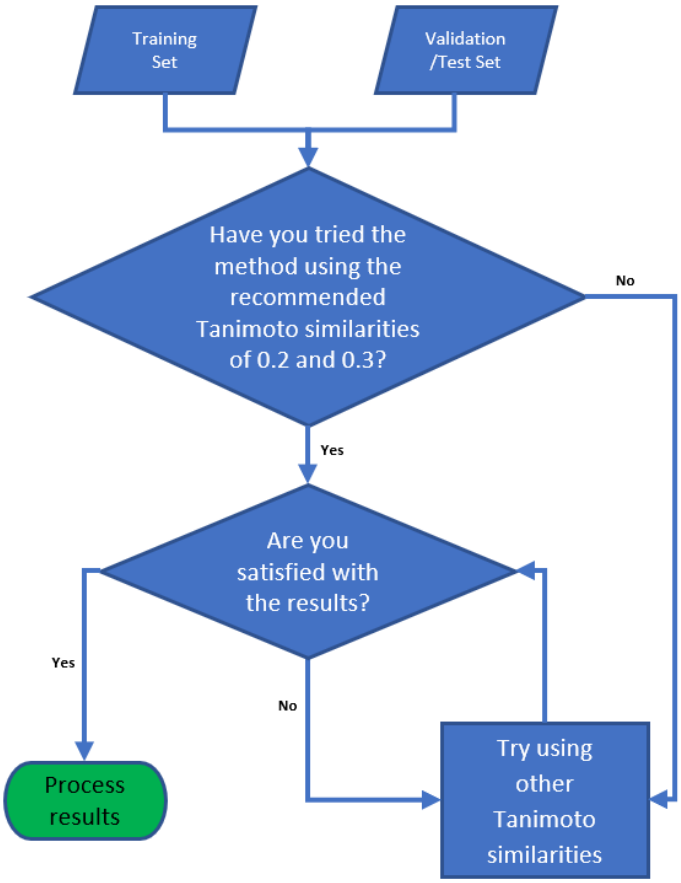


Figure A10: Workflow outlining how one should determine and apply the Tanimoto similarity given two datasets. If there are three datasets *i.e.* training, validation, and test sets, one should determine the similarity between the training and validation set, before determining the similarity between the training and test set. One should not assume that the data will be suitable for machine learning even if the data is taken from the same database.

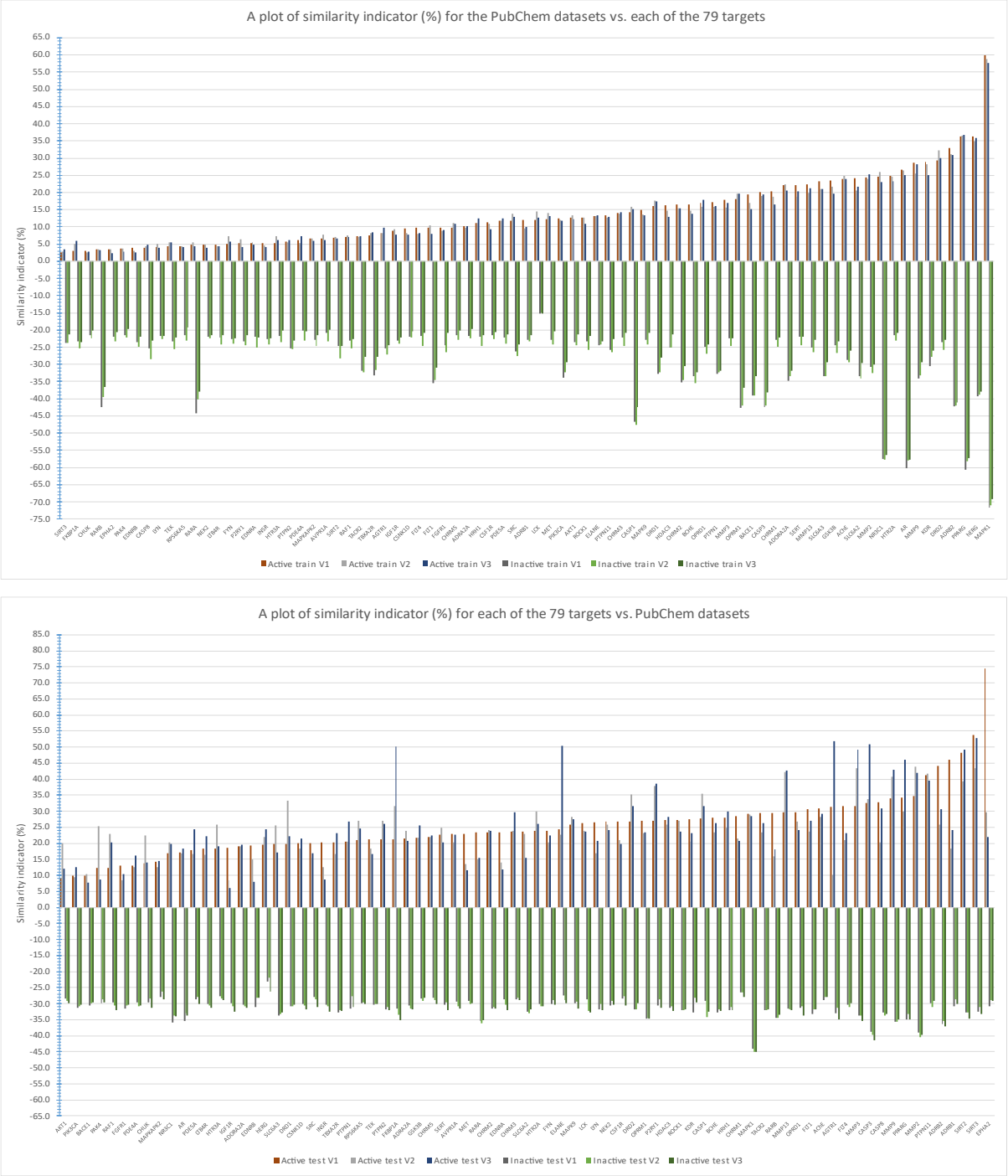


Figure A11: Plots showing the results of the similarity method when using it to compare randomly extracted datasets with a size of 1000 molecules from PubChem with the 79 human targets. Top: Results for PubChem datasets vs. each of the 79 targets. Bottom: Results for each of the each of the 79 targets vs. PubChem datasets. The positive bars represent comparison with target active data (data where all the target molecules are active) and the negative bars represent comparison with target inactive data (data where all the target molecules are inactive). A Tanimoto similarity threshold of 0.3 was used. The similarity indicator represents the proportion of molecules in the first dataset that are similar to the second dataset.

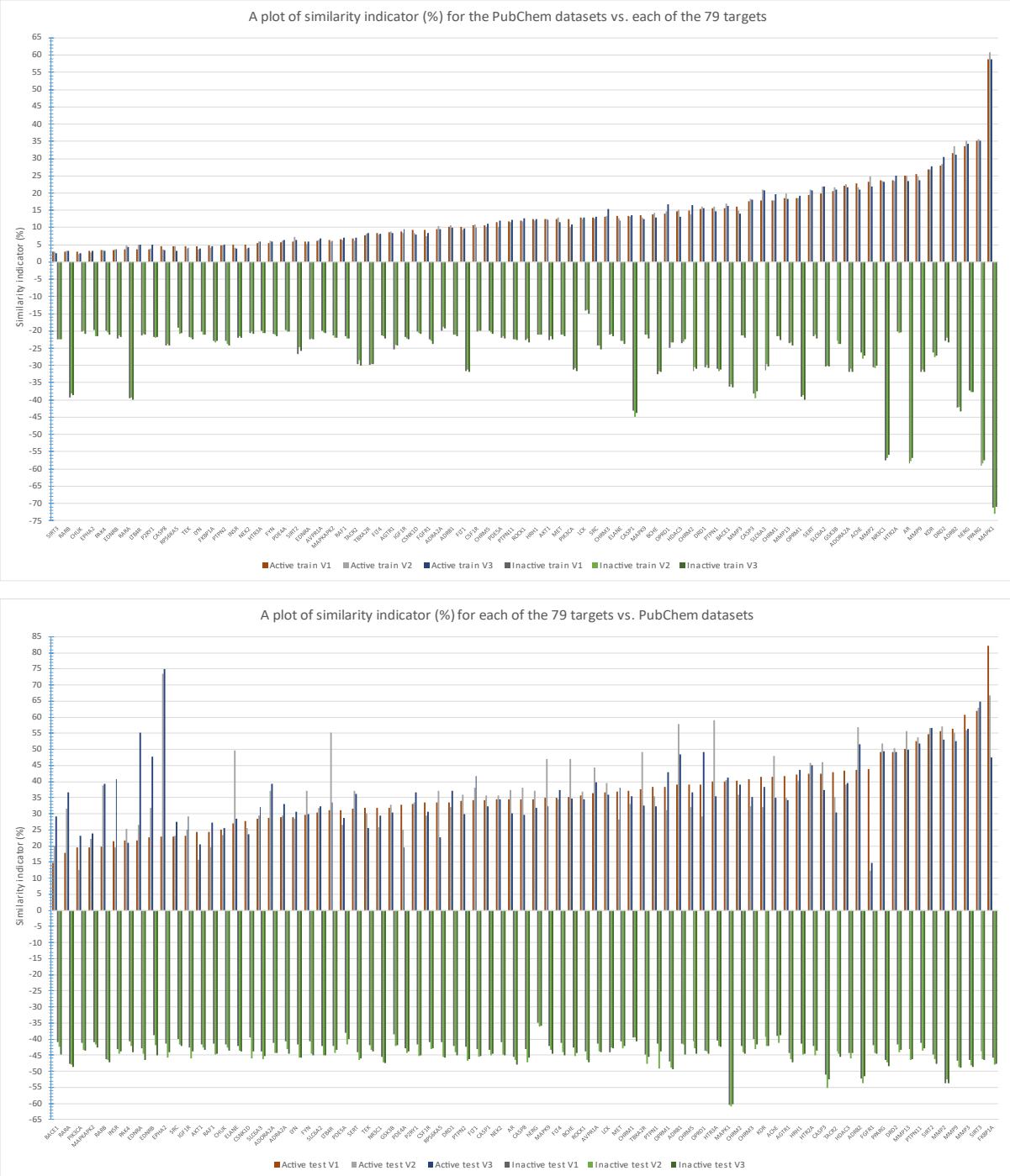


Figure A12: Plots showing the results of the similarity method when using it to compare randomly extracted datasets with a size of 2000 molecules from PubChem with the 79 human targets. Top: Results for PubChem datasets vs. each of the 79 targets. Bottom: Results for each of the each of the 79 targets vs. PubChem datasets. The positive bars represent comparison with target active data (data where all the target molecules are active) and the negative bars represent comparison with target inactive data (data where all the target molecules are inactive). A Tanimoto similarity threshold of 0.3 was used. The similarity indicator represents the proportion of molecules in the first dataset that are similar to the second dataset.

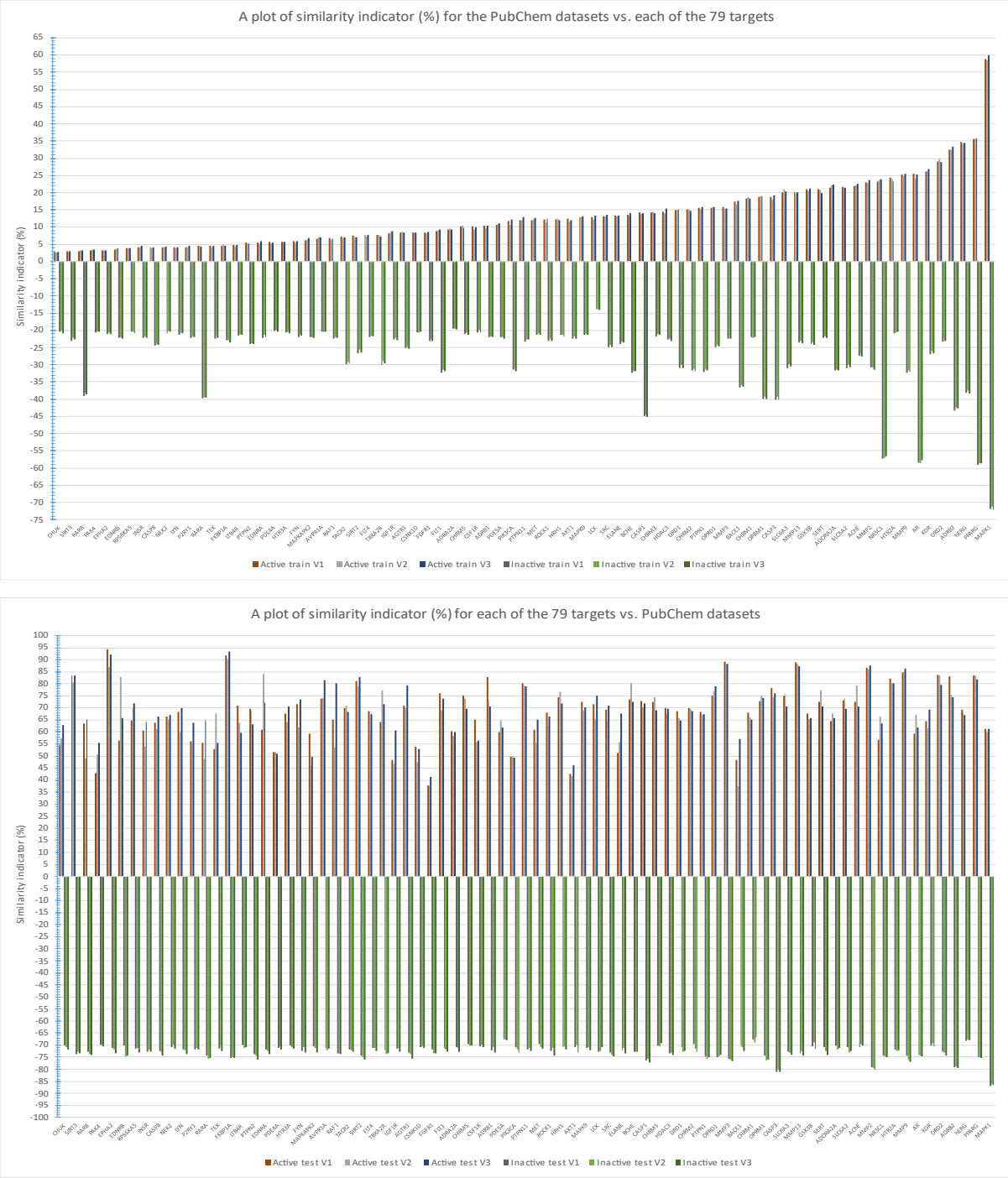


Figure A13: Plots showing the results of the similarity method when using it to compare randomly extracted datasets with a size of 10000 molecules from PubChem with the 79 human targets. Top: Results for PubChem datasets vs. each of the 79 targets. Bottom: Results for each of the each of the 79 targets vs. PubChem datasets. The positive bars represent comparison with target active data (data where all the target molecules are active) and the negative bars represent comparison with target inactive data (data where all the target molecules are inactive). A Tanimoto similarity threshold of 0.3 was used. The similarity indicator represents the proportion of molecules in the first dataset that are similar to the second dataset.

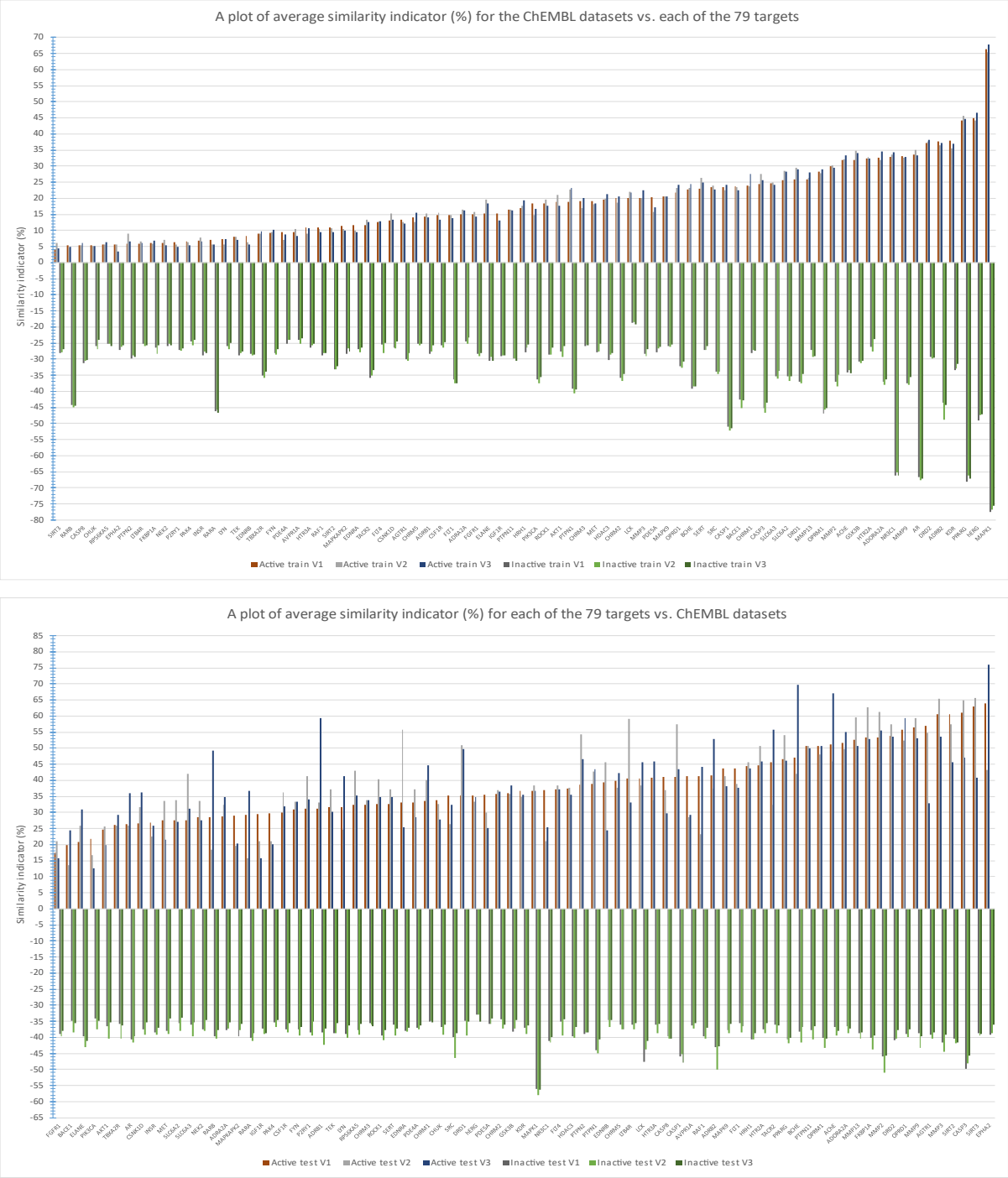


Figure A14: Plots showing the results of the similarity method when using it to compare randomly extracted datasets with a size of 1000 molecules from ChEMBL with the 79 human targets. Top: Results for ChEMBL datasets vs. each of the 79 targets. Bottom: Results for each of the each of the 79 targets vs. ChEMBL datasets. The positive bars represent comparison with target active data (data where all the target molecules are active) and the negative bars represent comparison with target inactive data (data where all the target molecules are

inactive). A Tanimoto similarity threshold of 0.3 was used. The similarity indicator represents the proportion of molecules in the first dataset that are similar to the second dataset.

Figure A15: Plots showing the results of the similarity method when using it to compare randomly extracted datasets with a size of 10000 molecules from ChEMBL with the 79 human targets. Top: Results for ChEMBL datasets vs. each of the 79 targets. Bottom: Results for each of the each of the 79 targets vs. ChEMBL datasets. The positive bars represent comparison with target active data (data where all the target molecules are active) and the negative bars

represent comparison with target inactive data (data where all the target molecules are inactive). A Tanimoto similarity threshold of 0.3 was used. The similarity indicator represents the proportion of molecules in the first dataset that are similar to the second dataset.

7.3: Chapter 4 Appendices

Table A16: Full results for the test performance using the consensus approach with varying consensus thresholds for developmental toxicity and reproductive toxicity separately

Toxicity type	No. of models used for majority prediction	SE (%)	SP (%)	Accuracy (%)	MCC	TP	FP	FN	TN
Developmental	13	70.7	63.1	67.3	0.338	171	73	71	125
	14	70.7	63.6	67.5	0.343	169	71	70	124
	15	70.7	63.4	67.4	0.341	164	71	68	123
	16	72.9	63.8	68.8	0.369	164	68	61	120
	17	72.9	63.4	68.6	0.364	161	67	60	116
	18	72.9	64.4	69.1	0.374	159	62	59	112
	19	73.0	65.3	69.6	0.383	157	59	58	111
	20	73.6	65.9	70.2	0.394	153	56	55	108
	21	74.5	65.8	70.6	0.404	146	54	50	104
	22	76.8	66.4	72.1	0.436	136	50	41	99
	23	78.6	68.3	74.0	0.471	121	39	33	84
	24	81.1	72.1	77.1	0.535	103	29	24	75
Reproductive	13	77.8	76.2	77.0	0.540	130	36	37	115
	14	77.8	77.0	77.5	0.548	130	34	37	114
	15	78.2	77.9	78.1	0.560	129	32	36	113
	16	78.0	77.9	78.0	0.559	128	32	36	113
	17	79.0	77.8	78.4	0.568	128	32	34	112
	18	78.8	78.0	78.4	0.567	126	31	34	110
	19	79.6	77.8	78.8	0.573	125	30	32	105
	20	80.1	79.7	79.9	0.596	125	26	31	102

	21	82.1	80.5	81.4	0.625	124	24	27	99
	22	83.0	79.5	81.4	0.624	122	24	25	93
	23	85.8	80.7	83.6	0.666	121	21	20	88
	24	87.0	82.8	85.3	0.697	114	16	17	77

Table A17: Results of best-performing models obtained by applying transfer learning of developmental toxicity models to human target training datasets

No.	Target	ACC (%)	SE (%)	SP (%)	MCC
1	AChE	63.6 ± 1.9	82.0 ± 4.5	38.5 ± 2.1	0.231 ± 0.041
2	ADORA2A	73.3 ± 2.3	84.4 ± 3.9	52.6 ± 1.3	0.393 ± 0.045
3	ADRA2A	70.2 ± 0.4	86.6 ± 0.9	56.3 ± 0.7	0.444 ± 0.009
4	ADRB1	74.6 ± 1.6	92.4 ± 3.8	53.4 ± 2.1	0.506 ± 0.04
5	ADRB2	63.2 ± 0.3	85.5 ± 3.1	41.5 ± 2.4	0.3 ± 0.016
6	AGTR1	68.3 ± 1.6	90.2 ± 3.3	53.3 ± 1.7	0.447 ± 0.039
7	AKT1	80.1 ± 1.2	93.5 ± 2.6	49.3 ± 2.2	0.499 ± 0.029
8	AR	54.5 ± 0.5	79.3 ± 3.0	45.5 ± 1.5	0.225 ± 0.017
9	AVPR1A	67.1 ± 2.0	86.1 ± 6.4	56.0 ± 1.3	0.415 ± 0.061
10	BACE1	77.6 ± 1.5	92.1 ± 2.9	44.8 ± 2.1	0.435 ± 0.035
11	BCHE	59.6 ± 1.7	77.8 ± 4.9	47.9 ± 0.8	0.259 ± 0.049
12	CASP1	47.2 ± 0.8	75.2 ± 9.4	35.5 ± 4.6	0.107 ± 0.051
13	CASP3	61.1 ± 2.0	77.5 ± 9.7	50.4 ± 3.2	0.282 ± 0.074
14	CASP8	57.2 ± 2.1	65.9 ± 12.8	54.6 ± 1.5	0.172 ± 0.098
15	CHRM1	70.4 ± 2.0	84.2 ± 4.4	48.5 ± 2.0	0.355 ± 0.04
16	CHRM2	65.5 ± 0.8	79.6 ± 1.8	54.0 ± 1.4	0.343 ± 0.018
17	CHRM3	69.5 ± 1.6	81.8 ± 3.5	51.9 ± 1.8	0.357 ± 0.032
18	CHRM5	67.4 ± 0.9	90.5 ± 1.0	52.8 ± 1.7	0.439 ± 0.013
19	CHUK	64.3 ± 0.9	92.2 ± 3.5	56.4 ± 1.5	0.405 ± 0.024
20	CSF1R	73.4 ± 3.2	87.1 ± 6.0	56.0 ± 0.7	0.462 ± 0.07
21	CSNK1D	67.3 ± 1.8	82.7 ± 5.5	56.9 ± 1.3	0.395 ± 0.051
22	DRD1	69.0 ± 1.0	90.1 ± 1.8	54.4 ± 1.6	0.456 ± 0.021
23	DRD2	88.1 ± 1.3	95.8 ± 1.9	49.8 ± 1.7	0.53 ± 0.037
24	EDNRA	67.2 ± 1.8	77.2 ± 4.2	55.9 ± 1.6	0.341 ± 0.039

25	EDNRB	64.6 ± 1.8	80.7 ± 4.7	54.2 ± 1.2	0.347 ± 0.047
26	ELANE	70.2 ± 1.8	82.0 ± 4.4	51.8 ± 4.2	0.357 ± 0.037
27	EPHA2	66.8 ± 0.9	92.1 ± 4.6	54.9 ± 1.6	0.448 ± 0.036
28	FGFR1	76.0 ± 2.4	90.3 ± 4.7	50.7 ± 2.1	0.461 ± 0.05
29	FKBP1A	64.1 ± 1.3	94.1 ± 3.5	53.8 ± 2.1	0.424 ± 0.027
30	FLT1	63.2 ± 2.1	80.7 ± 7.5	54.6 ± 1.4	0.336 ± 0.067
31	FLT4	66.9 ± 1.9	86.1 ± 6.8	55.7 ± 1.5	0.413 ± 0.06
32	FYN	64.2 ± 1.4	81.4 ± 7.1	57.3 ± 1.9	0.352 ± 0.054
33	GSK3B	71.9 ± 3.1	82.6 ± 5.0	49.6 ± 1.0	0.339 ± 0.053
34	HDAC3	58.5 ± 4.6	62.8 ± 10.7	54.6 ± 1.0	0.177 ± 0.103
35	HRH1	75.0 ± 1.0	93.4 ± 2.3	53.3 ± 1.3	0.519 ± 0.026
36	HTR2A	83.6 ± 1.3	91.5 ± 1.7	55.1 ± 1.1	0.494 ± 0.03
37	HTR3A	61.9 ± 0.9	79.7 ± 3.7	54.0 ± 1.1	0.314 ± 0.031
38	IGF1R	80.1 ± 3.0	91.1 ± 5.0	54.8 ± 1.8	0.508 ± 0.063
39	INSR	65.0 ± 2.6	79.1 ± 7.2	53.9 ± 1.3	0.337 ± 0.066
40	KCNH2	65.5 ± 0.5	89.3 ± 2.2	29.9 ± 2.4	0.244 ± 0.014
41	KDR	74.9 ± 3.7	80.0 ± 4.9	50.2 ± 2.1	0.264 ± 0.039
42	LCK	69.3 ± 4.2	76.9 ± 6.2	45.1 ± 2.8	0.212 ± 0.049
43	LTB4R	56.4 ± 1.6	54.4 ± 7.4	57.1 ± 1.4	0.1 ± 0.058
44	LYN	65.0 ± 1.3	83.5 ± 6.9	57.0 ± 1.4	0.375 ± 0.054
45	MAPK1	53.9 ± 1.3	72.6 ± 3.6	43.3 ± 3.8	0.159 ± 0.013
46	MAPK9	68.5 ± 2.1	79.9 ± 4.7	55.8 ± 1.1	0.37 ± 0.045
47	MAPKAPK2	66.7 ± 1.1	90.0 ± 2.6	50.4 ± 1.9	0.421 ± 0.026
48	MET	76.6 ± 6.0	84.8 ± 8.6	55.0 ± 1.4	0.415 ± 0.101
49	MMP13	70.8 ± 2.5	78.9 ± 4.5	52.6 ± 2.0	0.317 ± 0.036
50	MMP2	68.9 ± 1.3	83.9 ± 2.9	43.0 ± 2.8	0.296 ± 0.026
51	MMP3	72.9 ± 2.0	86.6 ± 3.4	49.7 ± 1.2	0.397 ± 0.045
52	MMP9	67.6 ± 1.3	78.4 ± 2.4	52.6 ± 1.3	0.321 ± 0.027
53	NEK2	61.6 ± 1.3	80.5 ± 9.2	56.2 ± 1.7	0.306 ± 0.066
54	NR3C1	54.9 ± 0.9	76.2 ± 3.3	45.6 ± 1.3	0.206 ± 0.028
55	OPRD1	75.3 ± 4.1	85.8 ± 6.4	49.3 ± 2.5	0.378 ± 0.079
56	OPRM1	72.6 ± 2.9	87.9 ± 5.4	49.0 ± 1.5	0.412 ± 0.065

57	P2RY1	60.6 ± 1.3	69.7 ± 4.2	55.8 ± 1.1	0.244 ± 0.037
58	PAK4	63.1 ± 0.5	86.9 ± 5.6	54.6 ± 2.1	0.369 ± 0.034
59	PDE4A	63.6 ± 1.6	76.2 ± 5.8	55.9 ± 1.3	0.315 ± 0.049
60	PDE5A	65.0 ± 1.2	73.7 ± 2.6	53.2 ± 1.0	0.275 ± 0.023
61	PIK3CA	73.1 ± 5.2	81.9 ± 8.1	53.1 ± 1.9	0.362 ± 0.083
62	PPARG	52.6 ± 2.1	70.6 ± 6.2	41.9 ± 1.8	0.126 ± 0.06
63	PTPN1	57.0 ± 1.2	75.2 ± 6.7	44.4 ± 3.6	0.202 ± 0.043
64	PTPN11	57.3 ± 1.1	74.3 ± 6.9	52.3 ± 1.7	0.223 ± 0.05
65	PTPN2	57.8 ± 1.0	70.5 ± 8.1	54.2 ± 2.1	0.206 ± 0.054
66	RAF1	69.4 ± 2.2	81.3 ± 4.9	54.6 ± 1.4	0.376 ± 0.046
67	RARA	59.2 ± 1.3	72.6 ± 3.3	57.9 ± 1.6	0.178 ± 0.014
68	RARB	61.6 ± 1.2	74.1 ± 2.7	60.5 ± 1.5	0.193 ± 0.011
69	ROCK1	71.6 ± 1.8	86.3 ± 3.9	54.5 ± 0.9	0.435 ± 0.042
70	RPS6KA5	61.4 ± 1.0	84.3 ± 8.9	56.5 ± 1.2	0.312 ± 0.061
71	SIRT2	56.5 ± 1.2	75.5 ± 10.7	51.3 ± 2.2	0.223 ± 0.075
72	SIRT3	58.7 ± 1.0	79.8 ± 14.7	55.9 ± 0.9	0.23 ± 0.089
73	SLC6A2	76.5 ± 1.0	91.2 ± 2.6	54.0 ± 1.6	0.501 ± 0.025
74	SLC6A3	75.2 ± 0.6	91.7 ± 1.7	53.9 ± 1.5	0.503 ± 0.016
75	SLC6A4	85.4 ± 1.5	94.6 ± 2.1	52.3 ± 1.4	0.536 ± 0.041
76	SRC	72.1 ± 2.5	86.8 ± 5.4	45.2 ± 3.7	0.36 ± 0.047
77	TACR2	65.6 ± 1.4	85.5 ± 6.0	56.6 ± 1.4	0.395 ± 0.05
78	TBXA2R	64.2 ± 1.7	79.1 ± 5.6	56.1 ± 1.3	0.339 ± 0.05
79	TEK	61.9 ± 2.1	72.7 ± 5.7	54.4 ± 1.2	0.27 ± 0.053

Table A18: Results of best-performing models obtained by applying transfer learning of reproductive toxicity models to human target training datasets

No.	Target	ACC (%)	SE (%)	SP (%)	MCC
1	AChE	65.0 ± 0.6	93.0 ± 1.4	26.7 ± 2.8	0.271 ± 0.014
2	ADORA2A	79.9 ± 0.5	98.8 ± 0.2	44.5 ± 1.5	0.558 ± 0.012
3	ADRA2A	65.2 ± 1.1	96.5 ± 0.8	38.8 ± 1.9	0.42 ± 0.019
4	ADRB1	68.6 ± 0.8	97.0 ± 0.5	34.9 ± 1.7	0.417 ± 0.017
5	ADRB2	58.0 ± 0.6	96.1 ± 0.3	20.8 ± 1.0	0.256 ± 0.014

6	AGTR1	59.1 ± 1.2	96.4 ± 0.5	33.6 ± 2.2	0.36 ± 0.015
7	AKT1	79.4 ± 1.0	99.1 ± 0.2	34.0 ± 3.5	0.487 ± 0.027
8	AR	55.5 ± 1.2	87.7 ± 0.5	43.8 ± 1.7	0.292 ± 0.01
9	AVPR1A	59.2 ± 1.0	98.8 ± 0.2	36.3 ± 1.6	0.4 ± 0.013
10	BACE1	80.1 ± 0.7	98.8 ± 0.4	37.7 ± 2.9	0.511 ± 0.017
11	BCHE	60.0 ± 1.3	92.9 ± 1.3	38.7 ± 2.8	0.351 ± 0.012
12	CASP1	39.1 ± 0.5	97.0 ± 0.8	15.0 ± 0.9	0.171 ± 0.008
13	CASP3	52.5 ± 0.8	96.0 ± 0.4	24.2 ± 1.2	0.267 ± 0.015
14	CASP8	48.3 ± 2.1	95.0 ± 1.1	34.8 ± 2.8	0.277 ± 0.019
15	CHRM1	72.4 ± 0.8	96.5 ± 0.3	34.1 ± 2.0	0.413 ± 0.019
16	CHRM2	66.9 ± 0.6	94.4 ± 0.4	44.4 ± 0.9	0.436 ± 0.01
17	CHRM3	70.8 ± 0.4	95.9 ± 0.8	35.2 ± 1.2	0.408 ± 0.011
18	CHRM5	58.2 ± 0.9	94.5 ± 0.6	35.3 ± 1.6	0.341 ± 0.012
19	CHUK	50.4 ± 1.9	97.1 ± 1.3	37.2 ± 2.8	0.311 ± 0.009
20	CSF1R	71.6 ± 0.8	99.3 ± 0.2	36.6 ± 1.8	0.48 ± 0.016
21	CSNK1D	62.7 ± 1.6	97.7 ± 0.4	39.0 ± 2.8	0.421 ± 0.018
22	DRD1	66.5 ± 1.7	94.7 ± 0.9	46.9 ± 3.3	0.447 ± 0.02
23	DRD2	87.8 ± 0.4	98.7 ± 0.3	33.6 ± 3.2	0.482 ± 0.024
24	EDNRA	68.2 ± 0.5	98.9 ± 0.6	33.7 ± 1.0	0.439 ± 0.011
25	EDNRB	59.1 ± 1.6	99.4 ± 0.6	32.9 ± 3.0	0.394 ± 0.015
26	ELANE	68.8 ± 0.5	95.8 ± 0.3	26.6 ± 1.6	0.325 ± 0.013
27	EPHA2	56.6 ± 2.1	99.0 ± 0.4	36.9 ± 3.2	0.383 ± 0.019
28	FGFR1	75.1 ± 0.7	98.8 ± 0.3	33.1 ± 2.4	0.46 ± 0.015
29	FKBP1A	51.4 ± 1.5	98.6 ± 0.5	35.3 ± 2.1	0.334 ± 0.013
30	FLT1	63.0 ± 1.2	98.8 ± 0.2	45.3 ± 1.8	0.449 ± 0.014
31	FLT4	59.0 ± 1.2	98.3 ± 0.4	35.9 ± 1.8	0.391 ± 0.018
32	FYN	53.1 ± 2.1	95.0 ± 0.5	36.4 ± 3.1	0.318 ± 0.02
33	GSK3B	76.0 ± 0.4	98.3 ± 0.3	29.5 ± 1.2	0.422 ± 0.012
34	HDAC3	60.9 ± 0.6	90.4 ± 1.6	34.6 ± 2.0	0.298 ± 0.013
35	HRH1	68.4 ± 0.7	98.4 ± 0.5	33.3 ± 1.4	0.43 ± 0.015
36	HTR2A	84.7 ± 0.5	98.5 ± 0.4	34.7 ± 1.8	0.486 ± 0.023
37	HTR3A	53.2 ± 1.4	97.0 ± 0.4	33.6 ± 2.0	0.33 ± 0.015

38	IGF1R	79.4 ± 0.6	99.4 ± 0.2	33.7 ± 1.9	0.493 ± 0.019
39	INSR	63.2 ± 1.0	98.4 ± 0.3	35.2 ± 2.1	0.415 ± 0.013
40	KCNH2	64.5 ± 0.3	97.6 ± 0.2	15.0 ± 0.9	0.237 ± 0.007
41	KDR	87.0 ± 0.2	98.2 ± 0.2	32.8 ± 1.6	0.453 ± 0.012
42	LCK	78.9 ± 0.4	98.4 ± 0.7	17.2 ± 2.1	0.297 ± 0.019
43	LTB4R	52.0 ± 2.2	95.0 ± 1.2	38.0 ± 3.2	0.311 ± 0.017
44	LYN	55.9 ± 1.3	99.1 ± 0.3	37.2 ± 1.8	0.379 ± 0.014
45	MAPK1	43.8 ± 1.5	91.4 ± 1.8	17.1 ± 3.3	0.116 ± 0.012
46	MAPK9	70.3 ± 1.0	99.4 ± 0.1	37.9 ± 2.1	0.483 ± 0.018
47	MAPKAPK2	60.1 ± 0.9	98.7 ± 0.3	32.9 ± 1.8	0.391 ± 0.01
48	MET	80.8 ± 0.5	99.1 ± 0.0	33.1 ± 1.6	0.485 ± 0.015
49	MMP13	74.6 ± 1.3	92.6 ± 1.7	34.2 ± 1.5	0.341 ± 0.036
50	MMP2	67.3 ± 1.0	91.8 ± 1.4	25.0 ± 1.0	0.231 ± 0.028
51	MMP3	70.8 ± 1.3	93.1 ± 1.5	33.2 ± 1.9	0.341 ± 0.037
52	MMP9	70.0 ± 0.8	89.1 ± 1.3	43.1 ± 1.0	0.37 ± 0.019
53	NEK2	51.0 ± 2.1	96.4 ± 0.9	38.0 ± 3.0	0.311 ± 0.012
54	NR3C1	57.5 ± 1.0	86.7 ± 0.9	44.9 ± 1.7	0.303 ± 0.007
55	OPRD1	79.6 ± 0.6	99.0 ± 0.2	31.6 ± 2.4	0.466 ± 0.019
56	OPRM1	75.7 ± 0.9	98.6 ± 0.5	40.2 ± 3.2	0.509 ± 0.016
57	P2RY1	56.7 ± 1.3	97.6 ± 1.0	35.4 ± 1.9	0.366 ± 0.018
58	PAK4	51.8 ± 1.5	97.0 ± 0.2	35.7 ± 2.1	0.324 ± 0.014
59	PDE4A	61.8 ± 1.7	98.2 ± 0.7	39.5 ± 3.0	0.421 ± 0.019
60	PDE5A	70.3 ± 0.8	98.0 ± 0.3	32.9 ± 1.9	0.427 ± 0.015
61	PIK3CA	83.0 ± 0.5	99.6 ± 0.2	44.9 ± 1.8	0.593 ± 0.011
62	PPARG	57.6 ± 0.8	85.7 ± 0.9	40.9 ± 1.5	0.278 ± 0.01
63	PTPN1	51.4 ± 0.7	94.4 ± 0.6	21.5 ± 1.2	0.219 ± 0.013
64	PTPN11	44.6 ± 1.2	89.4 ± 0.9	31.5 ± 1.8	0.197 ± 0.006
65	PTPN2	46.9 ± 1.2	96.1 ± 0.7	33.0 ± 1.7	0.274 ± 0.006
66	RAF1	70.9 ± 0.7	98.9 ± 0.3	35.9 ± 1.8	0.464 ± 0.013
67	RARA	56.9 ± 2.7	86.3 ± 1.3	53.9 ± 3.1	0.234 ± 0.012
68	RARB	56.5 ± 1.6	95.0 ± 0.7	53.0 ± 1.8	0.264 ± 0.011
69	ROCK1	69.4 ± 1.2	99.0 ± 0.5	34.9 ± 3.0	0.454 ± 0.018

70	RPS6KA5	48.2 ± 2.6	97.8 ± 0.5	37.6 ± 3.2	0.292 ± 0.019
71	SIRT2	44.5 ± 0.8	91.8 ± 2.1	31.3 ± 1.0	0.217 ± 0.02
72	SIRT3	44.5 ± 2.6	94.3 ± 3.8	37.8 ± 3.4	0.219 ± 0.015
73	SLC6A2	75.4 ± 0.6	96.1 ± 0.5	43.6 ± 1.6	0.49 ± 0.013
74	SLC6A3	73.1 ± 0.3	95.7 ± 0.4	43.8 ± 0.9	0.477 ± 0.004
75	SLC6A4	83.7 ± 0.3	98.2 ± 0.5	31.6 ± 1.6	0.447 ± 0.015
76	SRC	73.4 ± 0.5	98.8 ± 0.3	27.2 ± 1.5	0.407 ± 0.014
77	TACR2	62.8 ± 2.3	96.1 ± 0.2	47.8 ± 3.4	0.429 ± 0.024
78	TBXA2R	62.7 ± 1.1	94.8 ± 0.5	45.2 ± 1.8	0.413 ± 0.012
79	TEK	61.0 ± 1.5	98.4 ± 1.3	35.4 ± 3.3	0.404 ± 0.011

Table A19: Results of best-performing models obtained by applying transfer learning of developmental toxicity models to human target combined (test + training) datasets

No.	Target	ACC (%)	SE (%)	SP (%)	MCC
1	AChE	63.4 ± 1.9	82.4 ± 4.5	38.1 ± 2.0	0.231 ± 0.041
2	ADORA2A	73.9 ± 2.2	84.9 ± 3.6	53.1 ± 1.3	0.404 ± 0.044
3	ADRA2A	69.7 ± 0.4	86.3 ± 0.9	55.8 ± 1.0	0.435 ± 0.007
4	ADRB1	73.6 ± 1.6	91.8 ± 3.7	52.3 ± 1.7	0.489 ± 0.041
5	ADRB2	63.2 ± 0.2	85.8 ± 3.3	41.3 ± 2.7	0.303 ± 0.015
6	AGTR1	68.4 ± 1.5	90.5 ± 3.5	53.3 ± 1.8	0.45 ± 0.038
7	AKT1	80.4 ± 1.2	93.6 ± 2.5	50.7 ± 2.2	0.512 ± 0.028
8	AR	54.4 ± 0.5	79.3 ± 3.1	45.4 ± 1.4	0.224 ± 0.019
9	AVPR1A	66.8 ± 2.1	85.8 ± 6.3	55.8 ± 1.4	0.409 ± 0.061
10	BACE1	77.8 ± 1.5	92.0 ± 2.9	45.1 ± 2.0	0.435 ± 0.035
11	BCHE	59.5 ± 1.8	78.0 ± 5.1	47.4 ± 0.9	0.258 ± 0.05
12	CASP1	47.1 ± 0.6	75.9 ± 9.5	34.8 ± 4.7	0.108 ± 0.051
13	CASP3	61.0 ± 2.0	77.8 ± 9.9	50.2 ± 3.1	0.283 ± 0.076
14	CASP8	57.1 ± 2.1	66.9 ± 12.7	54.2 ± 1.4	0.177 ± 0.098
15	CHRM1	70.6 ± 2.2	84.0 ± 4.6	48.9 ± 1.9	0.356 ± 0.044
16	CHRM2	65.5 ± 0.4	79.5 ± 1.7	54.1 ± 1.2	0.343 ± 0.011
17	CHRM3	69.0 ± 1.5	82.0 ± 3.5	51.0 ± 1.6	0.35 ± 0.031
18	CHRM5	67.4 ± 0.7	90.3 ± 1.3	53.1 ± 1.3	0.439 ± 0.013

19	CHUK	64.1 ± 0.7	92.6 ± 2.2	55.6 ± 1.3	0.407 ± 0.013
20	CSF1R	73.2 ± 3.1	86.6 ± 6.3	56.1 ± 1.2	0.457 ± 0.068
21	CSNK1D	66.7 ± 1.6	82.3 ± 5.1	56.0 ± 1.3	0.384 ± 0.045
22	DRD1	69.3 ± 0.8	90.2 ± 1.6	55.0 ± 1.3	0.461 ± 0.017
23	DRD2	88.2 ± 1.4	95.9 ± 1.9	49.6 ± 1.2	0.532 ± 0.042
24	EDNRA	66.1 ± 1.3	77.2 ± 3.3	53.6 ± 1.9	0.319 ± 0.027
25	EDNRB	64.2 ± 1.6	81.0 ± 5.0	53.2 ± 1.2	0.342 ± 0.047
26	ELANE	70.1 ± 1.9	81.4 ± 2.6	52.5 ± 3.9	0.355 ± 0.043
27	EPHA2	67.4 ± 0.8	92.3 ± 4.2	55.5 ± 1.4	0.457 ± 0.033
28	FGFR1	76.1 ± 2.4	90.2 ± 4.7	50.9 ± 2.2	0.462 ± 0.051
29	FKBP1A	63.2 ± 1.2	95.1 ± 2.4	52.0 ± 1.5	0.422 ± 0.025
30	FLT1	63.9 ± 2.2	80.8 ± 7.4	55.1 ± 1.2	0.346 ± 0.068
31	FLT4	67.8 ± 1.5	86.8 ± 5.6	56.0 ± 1.4	0.427 ± 0.049
32	FYN	62.9 ± 1.2	81.9 ± 6.5	55.5 ± 1.5	0.337 ± 0.05
33	GSK3B	72.0 ± 2.9	83.2 ± 4.9	49.2 ± 1.2	0.344 ± 0.052
34	HDAC3	58.6 ± 4.7	62.7 ± 11.1	54.9 ± 1.3	0.178 ± 0.104
35	HRH1	74.0 ± 0.9	93.3 ± 2.3	51.8 ± 1.2	0.504 ± 0.024
36	HTR2A	83.9 ± 1.2	91.6 ± 1.7	55.6 ± 1.2	0.501 ± 0.027
37	HTR3A	61.5 ± 1.0	77.1 ± 3.1	54.8 ± 1.4	0.294 ± 0.025
38	IGF1R	79.6 ± 2.9	91.0 ± 4.8	54.6 ± 1.7	0.505 ± 0.062
39	INSR	66.3 ± 2.6	79.8 ± 7.1	55.2 ± 1.4	0.359 ± 0.066
40	KCNH2	65.5 ± 0.6	89.4 ± 2.2	29.5 ± 2.3	0.242 ± 0.017
41	KDR	75.1 ± 3.7	80.3 ± 4.8	49.6 ± 1.9	0.262 ± 0.041
42	LCK	70.0 ± 4.1	77.4 ± 6.0	45.3 ± 3.0	0.217 ± 0.047
43	LTB4R	56.3 ± 1.7	55.7 ± 9.6	56.6 ± 1.4	0.107 ± 0.073
44	LYN	64.0 ± 1.5	82.6 ± 6.8	55.9 ± 1.2	0.356 ± 0.056
45	MAPK1	53.9 ± 1.4	72.6 ± 3.5	43.4 ± 3.9	0.158 ± 0.014
46	MAPK9	68.5 ± 2.1	79.9 ± 4.7	55.5 ± 1.2	0.368 ± 0.046
47	MAPKAPK2	67.7 ± 1.0	90.1 ± 2.6	51.7 ± 1.7	0.434 ± 0.025
48	MET	75.7 ± 5.7	84.5 ± 8.3	53.6 ± 1.4	0.398 ± 0.093
49	MMP13	71.0 ± 2.6	79.1 ± 4.6	53.5 ± 1.7	0.33 ± 0.04
50	MMP2	69.1 ± 1.4	83.7 ± 2.9	43.5 ± 2.6	0.298 ± 0.029

51	MMP3	73.3 ± 2.1	86.6 ± 3.8	50.7 ± 1.6	0.406 ± 0.048
52	MMP9	67.7 ± 1.3	78.2 ± 2.2	53.1 ± 1.3	0.324 ± 0.025
53	NEK2	61.1 ± 1.7	80.5 ± 10.8	55.7 ± 1.4	0.301 ± 0.081
54	NR3C1	55.2 ± 0.8	76.6 ± 3.2	45.9 ± 1.3	0.212 ± 0.026
55	OPRD1	75.3 ± 3.8	86.2 ± 5.9	48.4 ± 2.0	0.375 ± 0.077
56	OPRM1	72.8 ± 2.5	88.0 ± 4.5	49.0 ± 1.5	0.412 ± 0.056
57	P2RY1	60.4 ± 0.9	69.8 ± 3.7	55.6 ± 1.3	0.241 ± 0.029
58	PAK4	63.5 ± 0.6	87.4 ± 5.5	55.2 ± 1.5	0.375 ± 0.038
59	PDE4A	63.7 ± 2.0	76.0 ± 6.5	55.9 ± 1.5	0.314 ± 0.056
60	PDE5A	64.9 ± 1.3	73.9 ± 2.9	52.9 ± 1.5	0.275 ± 0.026
61	PIK3CA	72.9 ± 5.1	81.7 ± 7.9	53.1 ± 1.8	0.359 ± 0.082
62	PPARG	52.8 ± 2.2	70.6 ± 6.2	42.1 ± 1.7	0.128 ± 0.061
63	PTPN1	56.9 ± 1.2	75.2 ± 6.8	44.5 ± 3.6	0.203 ± 0.044
64	PTPN11	57.6 ± 0.8	75.9 ± 6.5	52.3 ± 1.5	0.237 ± 0.045
65	PTPN2	57.1 ± 0.9	69.4 ± 6.7	53.7 ± 2.0	0.191 ± 0.043
66	RAF1	70.1 ± 2.7	81.3 ± 5.7	56.2 ± 1.9	0.392 ± 0.057
67	RARA	61.3 ± 1.3	70.0 ± 2.2	60.3 ± 1.7	0.183 ± 0.006
68	RARB	58.8 ± 1.2	77.2 ± 3.7	57.2 ± 1.5	0.189 ± 0.017
69	ROCK1	71.8 ± 1.8	86.8 ± 4.2	54.5 ± 1.1	0.441 ± 0.044
70	RPS6KA5	61.5 ± 0.8	83.9 ± 8.6	56.6 ± 1.2	0.31 ± 0.058
71	SIRT2	56.5 ± 1.2	74.1 ± 10.4	51.5 ± 2.3	0.214 ± 0.072
72	SIRT3	59.4 ± 1.0	77.5 ± 11.8	56.8 ± 1.7	0.226 ± 0.069
73	SLC6A2	76.6 ± 0.9	91.3 ± 2.6	54.6 ± 1.6	0.507 ± 0.024
74	SLC6A3	75.5 ± 0.5	91.3 ± 1.8	54.7 ± 1.6	0.505 ± 0.014
75	SLC6A4	85.3 ± 1.6	94.4 ± 2.2	53.0 ± 1.5	0.538 ± 0.043
76	SRC	71.7 ± 2.7	86.2 ± 5.5	46.0 ± 3.5	0.359 ± 0.051
77	TACR2	65.7 ± 1.6	85.9 ± 5.8	56.5 ± 1.4	0.398 ± 0.052
78	TBXA2R	63.4 ± 1.7	78.3 ± 5.6	55.5 ± 1.1	0.324 ± 0.05
79	TEK	62.6 ± 1.8	74.4 ± 5.6	54.5 ± 1.3	0.288 ± 0.049

Table A20: Results of best-performing models obtained by applying transfer learning of reproductive toxicity models to human target combined (test + training) datasets

No.	Target	ACC (%)	SE (%)	SP (%)	MCC
1	AChE	64.2 ± 0.3	94.1 ± 1.3	24.5 ± 1.9	0.266 ± 0.009
2	ADORA2A	80.1 ± 0.6	98.7 ± 0.2	45.0 ± 1.7	0.561 ± 0.013
3	ADRA2A	63.2 ± 0.7	96.4 ± 0.6	35.6 ± 1.4	0.391 ± 0.01
4	ADRB1	68.2 ± 0.7	97.1 ± 0.5	34.5 ± 1.7	0.414 ± 0.015
5	ADRB2	57.9 ± 0.5	95.6 ± 0.4	21.5 ± 1.3	0.254 ± 0.008
6	AGTR1	59.4 ± 1.7	96.1 ± 1.0	34.4 ± 3.4	0.362 ± 0.017
7	AKT1	78.9 ± 0.8	99.1 ± 0.2	33.1 ± 3.2	0.479 ± 0.023
8	AR	55.6 ± 1.1	87.4 ± 0.4	44.1 ± 1.7	0.291 ± 0.01
9	AVPR1A	59.3 ± 1.2	98.8 ± 0.3	36.3 ± 1.9	0.4 ± 0.016
10	BACE1	80.1 ± 0.7	98.8 ± 0.3	37.1 ± 3.0	0.506 ± 0.019
11	BCHE	60.5 ± 1.5	93.1 ± 1.2	39.2 ± 3.1	0.358 ± 0.015
12	CASP1	38.9 ± 0.5	97.3 ± 0.9	13.9 ± 0.9	0.168 ± 0.011
13	CASP3	52.0 ± 0.6	95.7 ± 0.4	23.8 ± 1.0	0.259 ± 0.009
14	CASP8	49.1 ± 2.3	95.3 ± 0.9	35.6 ± 3.1	0.286 ± 0.019
15	CHRM1	72.0 ± 0.5	96.0 ± 0.3	33.1 ± 1.1	0.395 ± 0.012
16	CHRM2	66.9 ± 0.4	94.4 ± 0.4	44.8 ± 0.8	0.438 ± 0.007
17	CHRM3	69.8 ± 0.6	95.7 ± 0.9	34.2 ± 1.2	0.394 ± 0.016
18	CHRM5	59.0 ± 0.8	95.2 ± 0.6	36.3 ± 1.5	0.358 ± 0.009
19	CHUK	50.7 ± 2.1	97.6 ± 0.9	36.8 ± 3.0	0.318 ± 0.012
20	CSF1R	71.7 ± 1.1	99.0 ± 0.3	37.0 ± 3.0	0.477 ± 0.018
21	CSNK1D	62.2 ± 1.6	97.7 ± 0.7	37.6 ± 3.1	0.412 ± 0.017
22	DRD1	66.5 ± 1.7	94.5 ± 0.9	47.5 ± 3.4	0.447 ± 0.02
23	DRD2	87.9 ± 0.3	98.8 ± 0.3	33.3 ± 2.7	0.482 ± 0.019
24	EDNRA	67.7 ± 0.5	98.8 ± 0.5	32.9 ± 1.2	0.431 ± 0.012
25	EDNRB	59.1 ± 1.5	99.5 ± 0.5	32.6 ± 2.9	0.393 ± 0.015
26	ELANE	68.8 ± 0.3	96.0 ± 0.5	26.6 ± 0.9	0.329 ± 0.008
27	EPHA2	56.8 ± 2.0	98.7 ± 0.4	36.7 ± 3.0	0.381 ± 0.019
28	FGFR1	75.3 ± 0.5	99.2 ± 0.1	32.5 ± 1.7	0.465 ± 0.013
29	FKBP1A	51.5 ± 1.6	98.7 ± 0.5	34.9 ± 2.2	0.335 ± 0.015
30	FLT1	63.9 ± 1.1	98.9 ± 0.1	45.6 ± 1.7	0.46 ± 0.013
31	FLT4	60.4 ± 1.7	97.7 ± 0.7	37.1 ± 3.1	0.399 ± 0.018

32	FYN	52.7 ± 1.4	96.1 ± 0.4	35.7 ± 1.9	0.323 ± 0.016
33	GSK3B	75.6 ± 0.4	98.3 ± 0.3	29.5 ± 1.3	0.421 ± 0.013
34	HDAC3	61.2 ± 1.1	90.9 ± 2.1	33.8 ± 3.2	0.298 ± 0.023
35	HRH1	68.2 ± 0.8	98.0 ± 0.4	33.9 ± 1.7	0.424 ± 0.015
36	HTR2A	84.9 ± 0.5	98.6 ± 0.3	34.8 ± 2.0	0.491 ± 0.022
37	HTR3A	52.5 ± 1.3	96.2 ± 0.6	33.8 ± 2.0	0.318 ± 0.013
38	IGF1R	78.9 ± 0.7	99.1 ± 0.3	34.7 ± 2.8	0.492 ± 0.018
39	INSR	63.8 ± 0.9	98.6 ± 0.3	35.5 ± 1.8	0.422 ± 0.012
40	KCNH2	64.7 ± 0.2	98.1 ± 0.2	14.4 ± 0.8	0.242 ± 0.005
41	KDR	87.2 ± 0.2	98.3 ± 0.1	32.4 ± 1.7	0.454 ± 0.013
42	LCK	79.3 ± 0.5	98.2 ± 0.7	16.8 ± 2.2	0.284 ± 0.025
43	LTB4R	52.2 ± 2.0	94.9 ± 1.3	37.8 ± 3.0	0.312 ± 0.014
44	LYN	55.3 ± 1.3	98.7 ± 0.3	36.5 ± 1.9	0.369 ± 0.014
45	MAPK1	43.9 ± 1.4	91.4 ± 1.9	17.3 ± 3.3	0.12 ± 0.01
46	MAPK9	69.6 ± 0.9	99.5 ± 0.1	35.8 ± 1.9	0.468 ± 0.015
47	MAPKAPK2	60.1 ± 0.9	98.6 ± 0.3	32.6 ± 1.7	0.388 ± 0.009
48	MET	80.5 ± 0.6	99.0 ± 0.3	34.0 ± 2.8	0.488 ± 0.018
49	MMP13	74.0 ± 1.2	92.8 ± 1.5	33.6 ± 1.4	0.34 ± 0.033
50	MMP2	67.3 ± 1.1	91.7 ± 1.5	24.8 ± 1.0	0.226 ± 0.031
51	MMP3	70.7 ± 1.4	92.8 ± 1.6	33.3 ± 1.8	0.337 ± 0.038
52	MMP9	69.1 ± 0.8	88.4 ± 1.3	42.2 ± 0.9	0.352 ± 0.02
53	NEK2	50.2 ± 2.2	96.6 ± 0.8	37.2 ± 3.0	0.306 ± 0.014
54	NR3C1	57.8 ± 1.0	87.3 ± 0.9	45.0 ± 1.7	0.311 ± 0.008
55	OPRD1	79.5 ± 0.7	99.0 ± 0.2	31.4 ± 2.6	0.465 ± 0.021
56	OPRM1	76.1 ± 0.9	98.4 ± 0.6	41.1 ± 3.2	0.513 ± 0.015
57	P2RY1	56.9 ± 1.4	97.7 ± 0.8	36.2 ± 2.0	0.371 ± 0.019
58	PAK4	52.1 ± 2.1	97.2 ± 0.7	36.6 ± 3.0	0.329 ± 0.014
59	PDE4A	61.7 ± 1.7	98.3 ± 0.6	38.2 ± 3.0	0.418 ± 0.019
60	PDE5A	70.0 ± 0.7	98.2 ± 0.3	32.8 ± 1.7	0.429 ± 0.012
61	PIK3CA	82.8 ± 0.5	99.7 ± 0.1	44.4 ± 1.8	0.589 ± 0.011
62	PPARG	57.9 ± 0.8	85.5 ± 0.9	41.3 ± 1.6	0.28 ± 0.01
63	PTPN1	50.9 ± 0.5	94.6 ± 0.6	21.4 ± 0.9	0.22 ± 0.011

64	PTPN11	44.5 ± 1.4	89.4 ± 0.4	31.4 ± 1.9	0.197 ± 0.012
65	PTPN2	45.5 ± 1.2	95.3 ± 0.6	31.6 ± 1.7	0.255 ± 0.006
66	RAF1	71.2 ± 0.7	99.0 ± 0.3	36.6 ± 1.8	0.472 ± 0.013
67	RARA	56.9 ± 3.0	86.9 ± 1.1	53.6 ± 3.4	0.242 ± 0.016
68	RARB	56.3 ± 2.9	93.7 ± 1.7	53.0 ± 3.3	0.256 ± 0.011
69	ROCK1	69.7 ± 1.2	99.2 ± 0.5	35.6 ± 2.9	0.463 ± 0.018
70	RPS6KA5	48.3 ± 2.4	97.7 ± 0.6	37.7 ± 3.0	0.291 ± 0.015
71	SIRT2	44.0 ± 0.8	91.4 ± 2.0	30.7 ± 1.1	0.208 ± 0.018
72	SIRT3	44.0 ± 2.4	91.5 ± 4.9	37.3 ± 3.4	0.201 ± 0.021
73	SLC6A2	75.3 ± 0.5	96.1 ± 0.5	44.2 ± 1.5	0.493 ± 0.011
74	SLC6A3	73.2 ± 0.8	95.6 ± 0.3	43.8 ± 1.8	0.476 ± 0.015
75	SLC6A4	83.6 ± 0.4	98.1 ± 0.4	32.0 ± 1.6	0.448 ± 0.019
76	SRC	72.9 ± 0.5	98.8 ± 0.3	27.3 ± 1.3	0.406 ± 0.012
77	TACR2	63.8 ± 2.3	96.8 ± 0.3	48.7 ± 3.4	0.445 ± 0.024
78	TBXA2R	61.5 ± 1.1	94.5 ± 0.6	44.1 ± 1.8	0.398 ± 0.012
79	TEK	61.5 ± 1.4	98.4 ± 1.1	35.7 ± 3.0	0.408 ± 0.011

Figure A16: A plot showing the optimal number of clusters for k-means clustering for the misclassified developmental toxicity data in Table 23

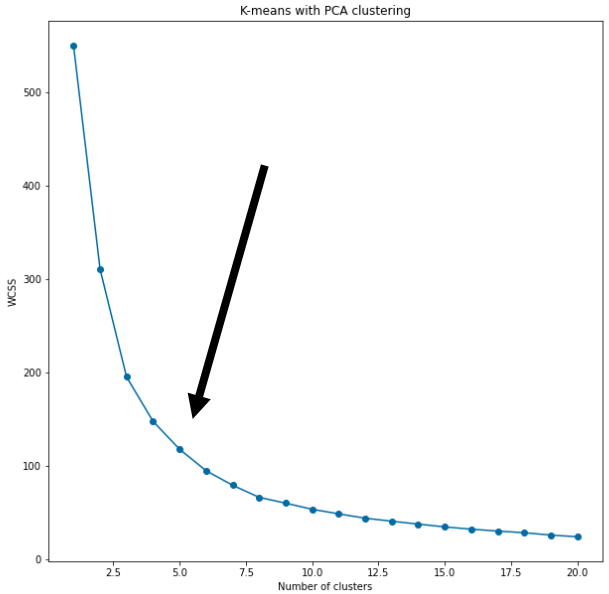


Figure A17: A plot showing the optimal number of clusters for k-means clustering for the misclassified reproductive toxicity data in Table 23