Towards More Reliable Drug Toxicity Prediction: An Ensemble Approach

by

Vladislav Yarovenko

Submitted to the Department of Computer Science in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science at

the

NAZARBAYEV UNIVERSITY

April 2024

Author .....

Department of Computer Science

05.04.2024

Certified by.....

Siamac Fazli

Associate Professor

Thesis Supervisor

Accepted by .....

Yelyzaveta Arkhangelsky Acting Dean, School of Engineering and Digital Sciences

Towards More Reliable Drug Toxicity Prediction: An Ensemble Approach by

Vladislav Yarovenko

Submitted to the Department of Computer Science on
05.04.2024, in partial fulfillment of the requirements for
the degree of
Master of Science in Computer Science

Abstract

The development of a single pharmaceutical drug is a time- and resource-consuming process with a high likelihood of rejection. In recent years, the cost-effectiveness of a single drug has decreased drastically, as the criteria for passing has become more rigorous. A huge fraction of attrition rates is caused by the toxicity of chemical compounds. Recent findings in Machine Learning (ML) have revolutionized the drug toxicity prediction field, developing many model architectures and data representations. The faced challenges are different ways of representing the molecules' chemical structure, as well as many different toxicity types. This study proposes a novel drug toxicity prediction framework. It uses several classification models, based on different data representations and different ways of combining their features. The evaluation of six different datasets with different toxicity types shows that choosing majority voting across all models can improve the ROC AUC score and accuracy. Using a single classification model to combine these datasets demonstrates that it is possible to achieve 84% accuracy on data with various toxicity types. The findings of this research provide insights into the application of ML in pharmaceutical research. Improving current methods of toxicity assessment can have a positive effect on the efficiency and cost-effectiveness of drug development.

Thesis Supervisor: Siamac Fazli
Title: Associate Professor

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The process of developing a single drug is very expencive, taking many years and billions of dollars. The efficiency of this process has been significantly declining over the decades, with the number of drugs developed per billion dollars decreasing from 50 in the 1950s to less than one today [1]. However, even with immense costs, there is a very high chance of rejection for potential drugs. Out of 10,000 potential drugs, only one manages to pass all the tests [2]. These tests usually include but are not limited to, assessing Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) properties. In this case, toxicity can be defined as a "diverse array of adverse effects which are brought about through drug use at either therapeutic or non-therapeutic doses" [3]. Toxicity alone is reported to be responsible for an attrition rate of 33% [4], which indicates its pivotal role in drug development and makes it essential to test for.

Recent discoveries in the Machine Learning (ML) field are revolutionizing drug discovery approaches. ML algorithms have already been applied in various domains of drug development, including drug design, biomarker identification, drug-target affinity prediction, and property prediction [5]. With the increasing number of available datasets, it is also possible to predict the toxicity of a molecule based only on its structure. One of the most common chemical structure representations is the Simplified Molecular-Input Line-Entry System (SMILES) [6]. However, while SMILES is a very detailed algorithm that does not loose a lot of information about chemical compounds, it is usually not used for ML models. Instead, it can be converted into more manageable formats, such as kernel matrices [7], feature matrices [8], graph representations [9], etc. These transformations improve the efficiency and scalability of ML algorithms in drug toxicity prediction, potentially leading to breakthroughs in pharmaceutical research.

Recent advances in the Machine Learning field and the existing variety of chemical data representations allow for a great variety of classification model architectures. Previous works report using Support Vector Machines, Random Forests, K-nearest neighbors, Convolutional Neural Networks, Graph Neural Networks, etc. [10]. However,

the optimal choice of a model depends not only on its individual performance but also on its synergy with the selected data representation. Different combinations of chemical representation and prediction algorithms can positively influence classification results.

More recent approaches go one step further and utilize several classification models by combining their results to get more accurate predictions [11, 12]. The idea behind this approach is that representations based on different physicochemical properties of molecules can provide complementary information. By employing several classification models and combining their output, it is possible to surpass the performance of individual models in terms of accuracy. Figure 1-1 illustrates the complete approach of converting a set of molecules into several data representations, training a separate ML classification model for each of them, and combining their predictions.

The complications of toxicity prediction stem from the existence of many toxicity types. Gola et al. state that toxicity is a "multi-factorial event with a plethora of possible responses" and that toxic response may result from many dose- and time-dependent chemical events [13]. Currently, there are several toxicity prediction datasets, with each of them describing hepatotoxicity [14], cardiotoxicity [15], oral toxicity [16], among others. Works covering several toxicity types tend to build a separate ML model for each. At this point, the possibility of creating a unified prediction model for several toxicity types is unclear. Therefore, our major contributions are as follows:

1. Implement several data representations, which do not lose information and are suitable for building ML classification models.
2. Find the best-performing ML classification model for each data representation.
3. Develop an ensemble approach for selected models that results in the best classification results.
4. Analyze the possibility of classifying different toxicity types with a single model.

5. Create a classification framework for chemical compounds and study its performance on established datasets.

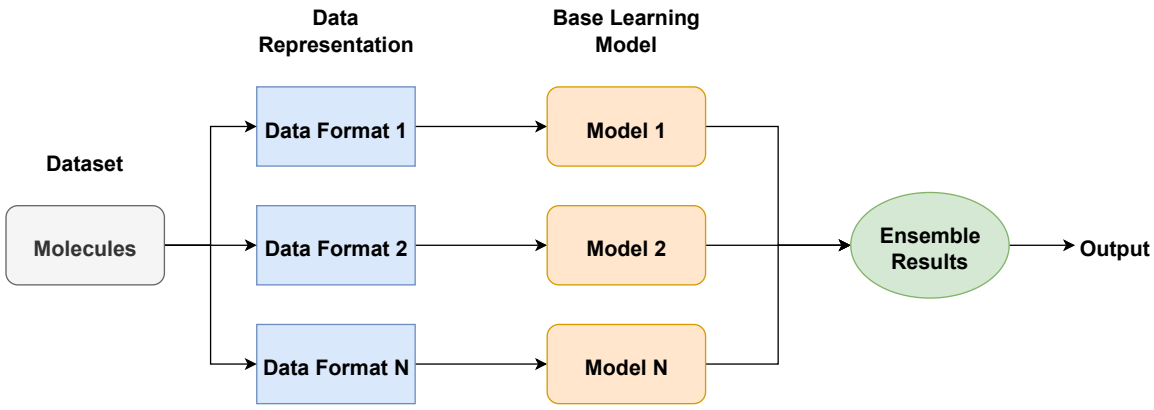


Figure 1-1: Baseline classifier architecture.

Section II discusses previous works on this topic and compares them to the proposed solution. Section III describes the complete architecture of the toxicity classifier and the methodology for its creation and evaluation. Section IV presents the obtained results and analyzes the performance of the created classifier. Section V summarizes all the findings, concludes this work, and proposes possible future work.

Chapter 2

Related work

2.1 Data Representations

The SMILES representation of molecules can be successfully converted to other data representations using various algorithms. Cao et al. [7] use the number of contiguous substrings in each compound to create a kernel matrix. Their model achieves an accuracy between 76% and 91%, depending on the used dataset. Authors in [17, 18] use a graph-based representation of compounds on several datasets, achieving ROC AUC score of

0.757 for the Tox21 dataset. The more common approach, however, is to convert SMILES into numerical data, such as feature matrices or molecular fingerprints, as proposed by Hirohara et al. [8]. Each symbol in SMILES is converted into a series of bits, where each bit corresponds to a specific property of a symbol: atom type, bond type, atom valence, etc. Chen, Cheong, and Siu implement the same approach, but the meaning of each allocated bit is slightly different [19]. The first approach achieves an AUC score of 0.813, while the second one shows an R-squared score of 0.619. Both used Tox21 as a training and testing dataset.

2.2 Individual Approaches

Different classification models are used in similar projects, with the model type being dependent on the SMILES representation type. Convolutional Neural Networks are often used for feature matrices and molecular fingerprints. Graph Neural Networks are used when the Graph representation is utilized. Other works also implement Support Vector Machines, Random Forests, etc. [7, 20].

In previous studies, the Tox21 Data Challenge is one of the most used datasets for developing toxicity prediction models. It contains a total of 8.5k compounds, split into groups that correspond to 12 targets. With this, the dataset has 12 binary classification tasks, where each compound is labeled as either toxic or non-toxic. The dataset, however, is not balanced, as non-toxic samples far outnumber the toxic ones. Other datasets, such as ClinTox and ToxB, show a better class ratio, with Clintox being perfectly balanced and ToxB having a ratio of 1:1.49. All three of these datasets can be obtained from MoleculeNet [18]. A large amount of datasets focus on a specific type of toxicity. The hERG Central dataset focuses on a "cardiac human Ether-à-go-go related gene (hERG) potassium channel" [15], which is an example of cardiotoxicity. Several datasets provide information about hepatotoxicity and liverinduced injuries [21, 22, 23]. Another dataset describes compounds obtained from Rat Acute Toxicity by Oral Exposure [16]. Table 2.1 illustrates several examples of toxicity prediction, along with their selected data representation and

ML models. While these approaches achieve good results, they all use only one data representation and model, and only a small number of works use several datasets.

Table 2.1: Examples of classification/regression approaches on toxicity datasets.

Source	Dataset(s)	Data Representation	Model	Results
[7]	DBPCAN NCTRER EPAFHM CPDBAS FDAMDD	SMILES-based strings (1D)	SVM	0.950 0.900 0.739 0.822 0.840
[8]	Tox21	Feature Matrix (2D)	2D CNN	0.877
[17]	Tox21	Graph (3D)	Graph CNN	0.757
[20]	Oral Acute	Fingerprints & Descriptors (1D)	RVM	0.679

2.3 Ensemble Approaches

The basis for the ensemble framework has been provided by Ryu et al. in 2020 [11]. The authors used three data representations and three corresponding classification models to get the final prediction on the hERG dataset. This was later adapted by Karim et al. in 2021 [12], where they used five data representations on several datasets. They do not use two-dimensional data, such as molecular feature matrices. On top of that, QTox experiments with several datasets separately instead of trying to merge them, while DeepHIT does not use several datasets at all [11]. Gupta and Rana also proposed an ensemble approach using three separate models [24]. However, all models were trained on the same data representation, and the work discussed only one dataset. Table 2.2 summarizes different ensemble approaches.

Table 2.2: Comparison of multi-model ensemble approaches.

Source	Data Representations	Models	Multiple Datasets?
ARE Ensemble [24]	Descriptors (1D)	Decision Tree Ada Boost SVM	No
DeepHIT [11]	Fingerprints (1D) Descriptors (1D)	DNN DNN	No

	Graph (3D)	Graph CNN	
QTox [12]	SMILES Vector (1D) Fingerprint Vector (1D) Fingerprints (1D) Descriptors (1D) Graph (3D)	1D CNN 1D CNN 2D CNN 2D CNN Graph CNN	Yes

Chapter 3

Methodology

3.1 Datasets

To test our framework for several types of toxicity, different datasets have to be selected. Using popular available frameworks, such as MoleculeNet and Therapeutics Data Commons (TDC), we made the choice based on data quality, completeness, and availability. The selection is also based on the prevalence of datasets in the existing literature. This allows for a more comprehensive comparison with prior research and analysis of results. Table 3.1 shows the final list of datasets that were included in this work.

Each toxicity dataset should provide the two following groups of data:

1. Structure Representation. A certain notation that represents the chemical structure of the molecule. Usually, datasets use SMILES notation due to its compactness, unambiguous representation, and information storage.

2. Toxicity Label. Binary (e.g., toxic/non-toxic) markers indicate the toxicity level of corresponding molecules. Labels should be assigned based on experimental or computational studies. One dataset can provide several types of labels, depending on tested assays and toxicity types.

Such datasets vary in size and toxicity type, ranging from several hundred to hundreds of thousands of molecules tested for cardiotoxicity, hepatotoxicity, cytotoxicity,

Table 3.1: Toxicity Datasets Used in This Work.

Dataset	Size	Obtained From	Label Type	Toxicity Type
Tox21	~8500	MoleculeNet	Binary	Nuclear receptor signals & stress response indicators
SIDER	1300	MoleculeNet	Binary	Various adverse drug reactions
ClinTox	1484	MoleculeNet	Binary	Various clinical trial drug toxicity types
hERG blockers	656	TDC	Binary	Cardiotoxicity
hERG Karim	13446	TDC	Binary	Cardiotoxicity
DILI	476	TDC	Binary	Hepatotoxicity
Combined	26554	Moleculenet & TDC	Binary	Various types

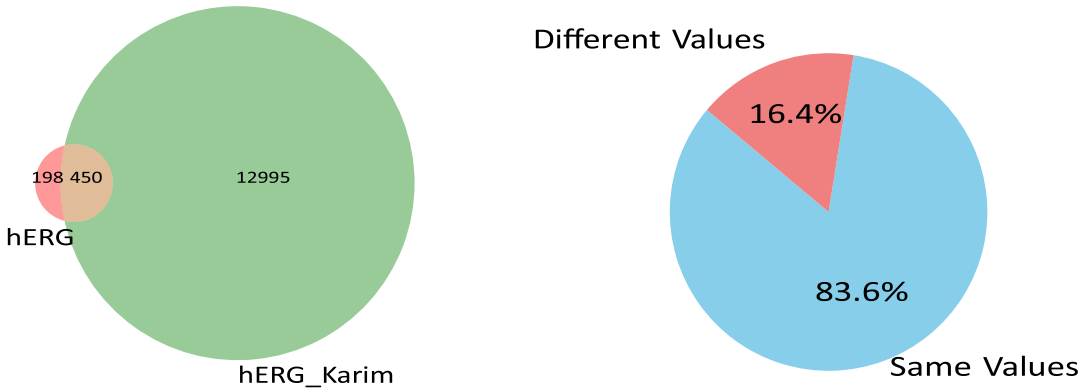
acute toxicity, etc.

ClinTox provides 1484 molecules that "were annotated as having failed for toxicity reasons during clinical trials" [25]. Each molecule has a binary toxic/non-toxic label, with the overwhelming majority of the molecules being toxic.

The Drug-Induced Liver Injury (DILI) dataset is gathered from the U.S. FDA's National Center for Toxicological Research [14]. It contains 475 molecules with binary labels, indicating whether they can cause a liver injury or not (hepatotoxicity).

hERG Central is a large database of more than 300,000 molecules [15]. It provides two regression tasks (inhibition at 1μM and 10μM concentration) and one classification task. Binary classification labels show whether a molecule blocks the Human ether-à-go-go related gene (hERG). The original hERG Central dataset is quite unbalanced, having only 5% of blocker molecules. Due to that, there exist more practical subsets by Karim et al. and Wang et al. that provide smaller but more balanced data [26, 27]. As shown in Figure 3-1a, they have 450 overlapping molecules, which is 70% of the hERG blockers subset. Additionally, out of these 450 molecules, 16.4% have different labels. These differences make it reasonable to test these subsets separately instead of combining them.

Tox21 was created in 2014 for a toxicity classification challenge as a collaboration between National Institute of Environmental Health Sciences (NIEHS) / National



(a) Overlap Between hERG Blockers (b) Labels of Overlapping Molecules. and hERG Karim et al.

Figure 3-1: Comparison of two hERG Subsets.

Toxicology Program (NTP), National Center for Advancing Translational Sciences (NCATS), U.S. Food and Drug Administration (FDA), and National Center for Computational Toxicology. It contains 12 binary labels: six nuclear receptor (NR) and six stress response (SR) pathway assays. Each assay has around 8500 molecules, with the majority present in several assays. However, the class distribution is not balanced, with every assay providing only around 7% toxic molecules.

The Side Effect Resource (SIDER) dataset contains 1300 molecules and adverse drug reactions to these molecules, grouped into 27 organ classes [28]. On average, each class provides balanced binary data labels.

All of the mentioned datasets can be obtained from popular drug toxicity frameworks, such as MoleculeNet [18] and Therapeutics Data Commons (TDC) [29]. They provide available versions of datasets with classification labels and a proper list of molecules stored as SMILES. Table 3.1 summarizes the information about the described datasets.

Class imbalance poses a significant challenge for Machine Learning problems, and drug toxicity prediction is no exception. As illustrated in Figure 3-3, only hERG Karim, DILI, and SIDER have nearly equal amounts of toxic and non-toxic molecules. In other datasets, such as hERG blockers, the imbalance is more pronounced, as it provides almost 69% of toxic molecules. ClinTox and Tox21 are even more unbalanced, having more than 90% of toxic

and non-toxic molecules, respectively. This can lead to biased model performance, with classifiers favoring the majority class and mispredicting the other.

Nonetheless, the Tox21 dataset was selected as the baseline for several reasons. First, it provides 12 assays, which are 12 different classification tasks. This allows models to be evaluated on one or several toxicity types without the need to use several datasets. Each classification task provides around 9000 molecules, leaving more than 1000 samples for training, validation, and testing sets. Due to this manageable size, evaluating models on Tox21 does not require a significant amount of time.

Finally, a combined dataset is constructed to consider the option of creating a unified classification model to predict molecules with different toxicity types. To do so, we use all of the molecules from SIDER, ClinTox, hERG blockers, hERG Karim, SIDER, Tox21, and DILI. Both SIDER and Tox21 provide several toxicity labels for each molecule. Because of the overlap and potential conflicts, different labels cannot be used in the combined dataset. Therefore, only the first classification is used for SIDER, while molecules and labels from the "NR-AR" assay are selected for the Tox21 dataset. Figure 3-2 shows that while some overlap between the datasets exists, it is relatively small compared to the size of the individual datasets. Therefore, removing duplicates does not have a huge effect on the overall size of the combined dataset. This results in a dataset with 26.5k molecules and many different types of toxicity.

3.2 Input Data Representations

Initially, all of the datasets provide samples stored as SMILES. This format stores most of the information about the molecule’s chemical structure while taking relatively little disk space. However, many related works on toxicity prediction propose converting SMILES into other data representations. This provides features with the required information and improves ML model compatibility.

After reviewing the literature, we were able to outline 3 prevalent data types, each having its own strengths and weaknesses:

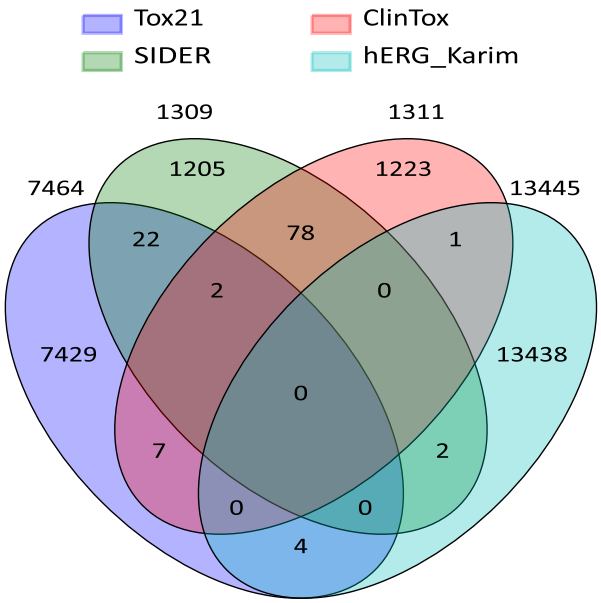


Figure 3-2: The Overlap of All Used Datasets.

1. Numerical data

2. Graph-based representation

3. Token-based representation

Numerical data tends to be the most flexible, providing a comprehensive range of features and suitable for a large range of model architectures. It is created by calculating the chemical properties of molecules, such as descriptors and fingerprints. This type of data is applicable for traditional Machine Learning algorithms, such as decision trees, support vector machines, and gradient boosting machines [30]. Deep learning models such as Deep Neural Networks (DNNs), one- and two-dimensional Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs) can also be considered.

Graph-based representations mostly specialize in capturing the structural dependencies of molecules and interactions within them. Usually, atoms are used as graph nodes and covalent bonds between them are used as edges. This structure closely resembles the actual structure of a molecule, accurately conveying its atomic posi-

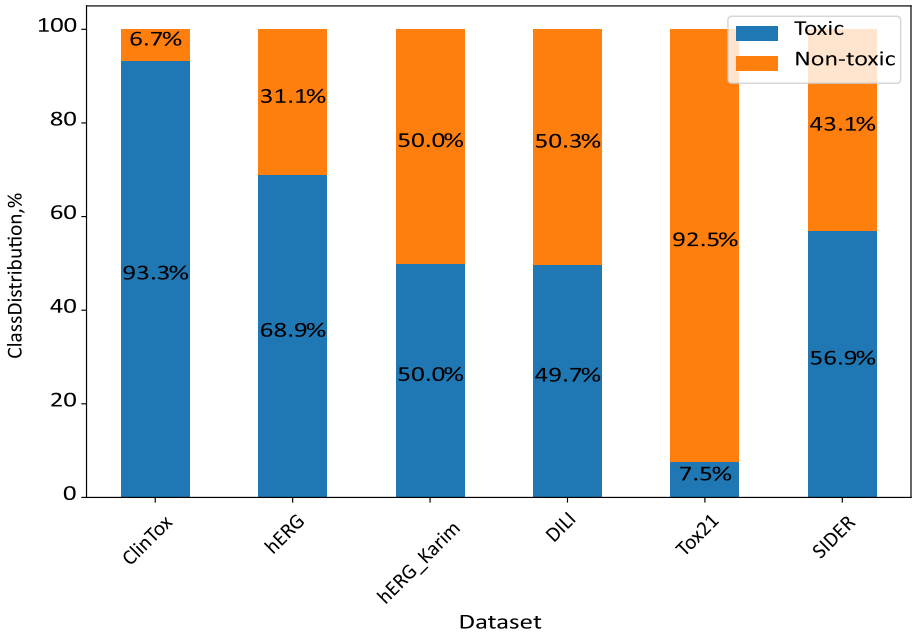


Figure 3-3: Class distribution of used datasets.

tions, bond types, and spatial orientation. Graph Neural Networks and their variants can process this type of data representation: Graph Convolutional Neural Networks, Graph Attention Networks, Graph Isomorphism Networks, etc.

The selection of tokens is based on the language models' recent advances in drug toxicity classification. SMILES is a formal language with tokens representing a molecule's chemical properties: atoms, bonds, rings, aromaticity, etc. Therefore, using SMILES effectively transforms drug toxicity classification into a Natural Language Processing task. Models like Bidirectional Encoder Representations from Transformers (BERT) and Long Short-Term Memory Networks (LSTM) have already been implemented for this task, outperforming some of the above-mentioned models [31, 32].

In our approach, we use all three of these data representations. Multiple numerical data representations are implemented, each providing a different subset of chemical properties, such as molecular fingerprints and chemical descriptors. One graph-based representation is employed to consider structural relationships and interactions within molecules. Finally, inspired by recent advancements in natural language processing, one

token-based data representation will capture intricate molecular characteristics and structural patterns.

3.2.1 Morgan Fingerprints

Circular or Morgan Fingerprints are a numerical data representation that encodes the chemical environment around each atom in a molecule within a defined radius. Morgan Fingerprints have previously been used in Quantitative Structure-Activity Relationship (QSAR) tasks, including toxicity classification. We used RDKit software to calculate fingerprints for selected datasets [33]. Using a bond radius of 2, fingerprints can be calculated as bit vectors with a length of 1024.

3.2.2 Chemical Descriptors

Two other numerical data representations calculate two-dimensional physicochemical descriptors using Mordred and RDKit packages, respectively. Mordred provides more than 700 descriptors about atom and bond counts, topological indices, molecular complexity indices, surface area properties, etc. On the other hand, RDKit descriptors calculate 43 descriptors related to molecular weights and sizes, surface areas, and aromaticity. The overlap in these descriptors is very small, and features that appear in both descriptor groups have different values. For example, both Mordred and RDKit calculate Lipinski's rule of five, which checks whether a drug violates a certain set of rules [34]. However, Mordred calculates one boolean descriptor, while RDKit provides two numerical ones, representing the amount of Ns/Os and N-H/O-H bonds respectively.

3.2.3 Graph-based Representation

In this data representation, atoms are used as nodes, and the connections between them are as edges. Chemical features related to them are calculated as well. This results in a data representation that captures both the molecule's topological and chemical

information. This work uses the DGL-LifeSci package for building graphbased representations of molecules [35].

3.2.4 Token-based representation

First, Open Babel software is used to calculate the molecule's atomic coordinates based on its SMILES representation [36]. This method is imperfect, as SMILES representation does not store this information, and coordinates must be approximated. This is done for every dataset except Tox21, which provides its own atomic coordinates. Next, an atomic pair distribution function (PDF) is calculated for each atom. According to Shermukhamedov et al., the PDF "represents the probability of finding an atom inside a sphere with a radius r centered at a selected atom" [37]. The resulting PDF vector is reduced using Principal Component Analysis and clustered with a K-means algorithm. After applying the PCA-KM algorithm, the final output is a set of tokens, each consisting of an atom from the molecule and its cluster.

3.3 Classification Model Selection

3.3.1 Random Forest Classifier

Random Forest (RF) is an ensemble Machine Learning algorithm that can be used for classification tasks. It constructs multiple Decision Trees using random subsets of training data and considers each output to calculate the final prediction (Figure 3-4). The ability to handle large amounts of data and the interpretability are the main advantages of this algorithm.

Random Forest has previously been implemented for Quantitative Structure-Activity Relationship (QSAR) tasks, including toxicity classification. A study by Wu et al. compared different model performances, contrasting simpler algorithms like Linear Regression, K-nearest neighbor, and Random Forest against more complex Deep Neural Networks [30]. Their results on Tox21 have shown that the difference between the results of simple and complex models is not significant. The best-performing architecture was Random Forest.

This study also covered different data representations, such as Morgan fingerprints and chemical descriptors from RDKit and Mordred. Given these findings, we also choose Random Forest as the model for numerical data representations. The architecture is implemented using the Scikit-learn Python package. The selected number of estimators was set to 100, and the class weight is "balanced." The last setting specifies that the model weights dynamically change based on the class balance of the provided dataset. This assigns higher weights to the minority class to mitigate the class imbalance of the datasets.

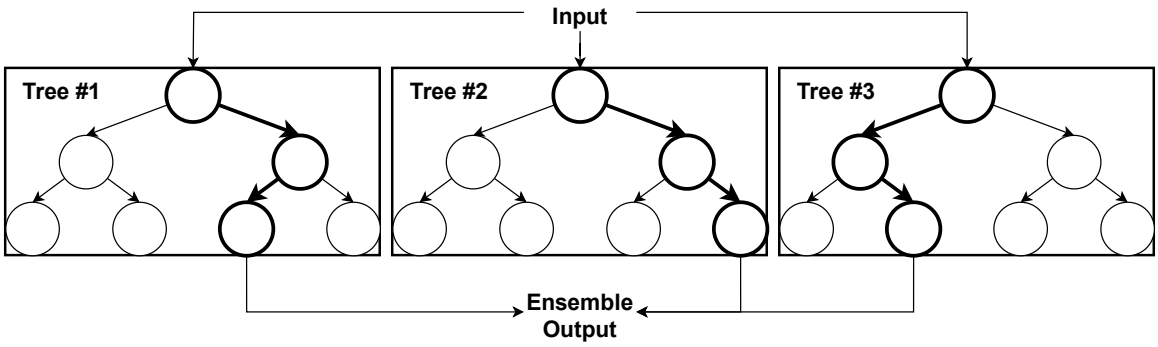


Figure 3-4: Representation of the Random Forest Classifier Algorithm.

3.3.2 Graph Isomorphism Network

A Graph Isomorphism Network (GIN) is a version of a Graph Neural Network, a Machine Learning model that uses graphs as input data. GIN introduces a more discriminative aggregator, which gathers information from neighboring nodes and allows them to exchange it [38]. GINs introduce an isomorphism test that tells if two graphs have the same structure with different permutations. Figure 3-5 illustrates two graphs with different node positions but identical connections, making them isomorphic. Combined with an additional Edge Prediction algorithm, this architecture has shown better performance in drug toxicity classification on the MoleculeNet datasets as compared to other graph-based models [39]. Based on these findings, we employ the same model for this work. This implementation is based on the DGL-LifeSci Python package, which

provides a framework for converting molecules into graphs and building graph-based models. The architecture and hyperparameters are unchanged from the original work.

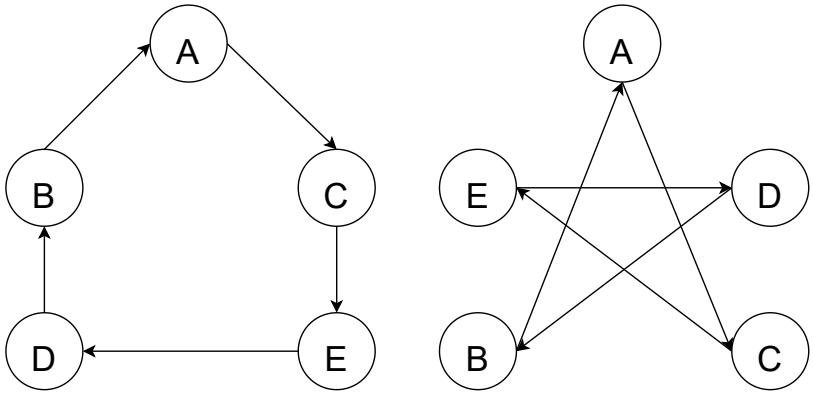


Figure 3-5: Representation of the Isomorphism in Graphs.

3.3.3 Deep Bidirectional Transformers

Bidirectional Encoder Representations from Transformers (BERT) is a Machine Learning language model that implements a transformer to capture both left and rightdirectional context from data [40]. During the training process, some of the tokens are masked to try to predict them based on the context. Treating SMILES symbols as tokens, Shermukhamedov et al. modified this model by adding a classification token and another layer to classify drug toxicity on Tox21, SIDER, and ClinTox datasets. Figure 3-6 illustrates the complete framework of the element Embeddings and Bidirectional Encoder Representations from Transformers (elEmBERT) model.

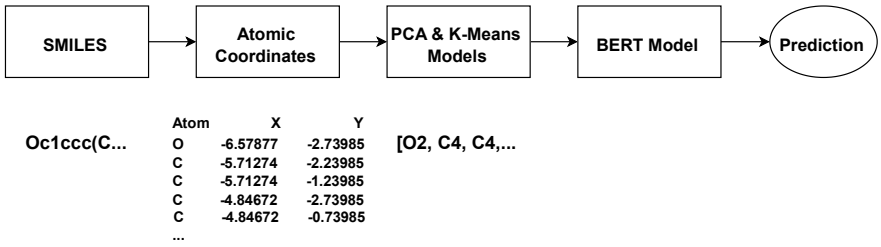


Figure 3-6: Complete architecture of the elEmBERT framework.

3.4 Ensemble Methods

Each model produces a binary output, 1 for toxic and 0 for non-toxic. In this work, we consider three possible methods to combine these results. The "Majority vote" outputs a result that was predicted by the majority of models. The molecule is toxic if at least half of the models (in our case, at least 3) predict that and is non-toxic otherwise (Equation 3.1).

$$pred_{majority} = \begin{cases} 1, & \text{if } \sum_{i=1}^{n} pred_i \geq \frac{n}{2} \\ 0, & \text{otherwise} \end{cases} \quad (3.1)$$

The "Minority vote" classifies a molecule as toxic if at least one model predicts the compound to be toxic (Equation 3.2.) This approach has been used previously, and it is meant to reduce the number of false positive predictions and improve precision. However, it also generally reduces overall accuracy. [11].

$$pred_{minority} = \begin{cases} \begin{cases} 1, & \text{if } \sum_{i=1}^{n} pred_i \geq 1 \end{cases} \\ 0, & \text{otherwise} \end{cases} \quad (3.2)$$

For the third ensemble approach, each model's performance is calculated for each dataset. The validation ROC AUC results are selected as the assigned weights for each model, and the weighted average $pred_{ensemble}$ is calculated, as shown in Equation 3.3. The final toxicity prediction is equal to 1 if $pred_{weighted}$ is greater than or equal to 0.5, and 0 otherwise. The combination of these results is the classifier framework's final part, illustrated in Figure 3-7.

$$pred_{weighted} = \begin{cases} 1 & \frac{\sum_{i=1}^{n} w_i \cdot pred_i}{\sum_{i=1}^{n} w_i} \geq 0.5 \quad , \quad \text{if} \\ 0, & \text{otherwise} \end{cases} \quad (3.3)$$

3.5 Evaluation Metrics

Common Machine Learning evaluation metrics are used to measure our model's performance, such as accuracy, precision, recall, and ROC AUC. Given the binary nature of the dataset labels (1/0 for toxic/non-toxic), it is possible to split all predictions into True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) ones. TP and FP are the number of toxic molecules classified as toxic and non-toxic, respectively. TN and FN are the number of non-toxic molecules classi-

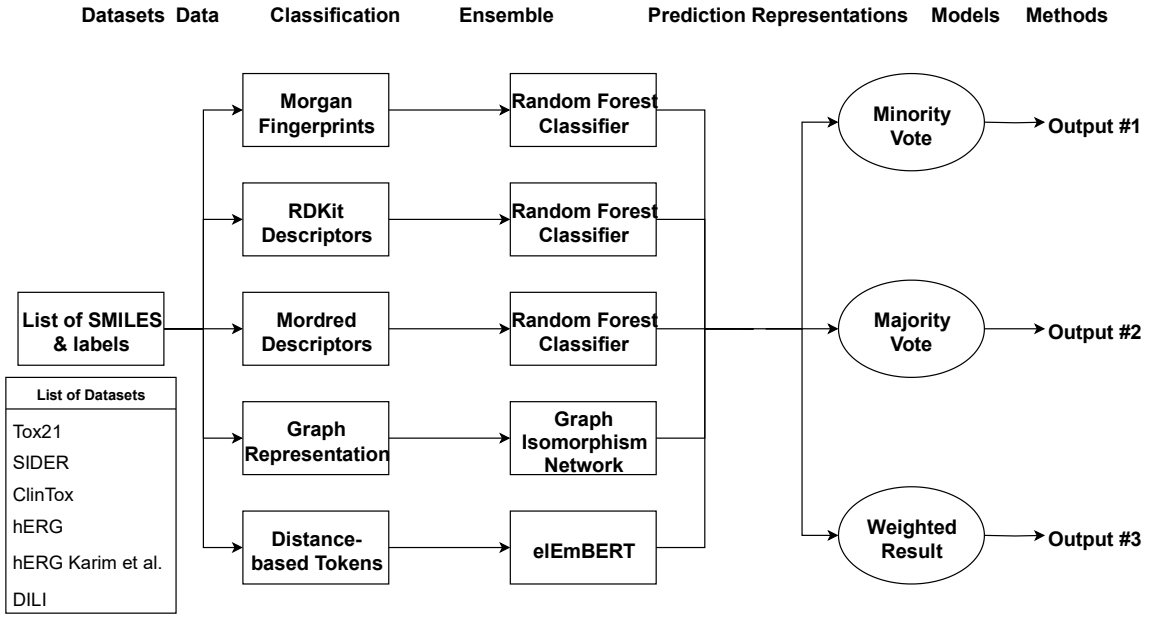


Figure 3-7: Framework for Toxicity Classifier.

fied as non-toxic and toxic. Accuracy is the ratio of correctly predicted samples to total samples (Equation 3.4). It measures the overall correctness of predictions and evaluates model performance across all classes.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.4}$$

Precision calculates the ratio of true positives to the sum of true and false positives (Equation 3.5). It reflects the model's ability to identify toxic molecules and reduce the mislabeling of non-toxic molecules as toxic.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3.5)$$

Recall is the ratio of true positives to the sum of true positives and false negatives (Equation 3.6). It measures the model's ability to identify all actual positive samples of the dataset. It is also known as True Positive Rate (TPR) and Sensitivity.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3.6)$$

Receiver Operating Characteristic Area Under the Curve (ROC AUC) shows the model's ability to distinguish between positive and negative samples across different classification thresholds. ROC-curve plots TPR against FPR, while AUC calculates the model's discriminatory power (Equation 3.7).

$$\text{ROC AUC} = \int_0^1 \text{TPR}(FPR)d(FPR) \quad (3.7)$$

All metrics are calculated for each dataset, including the combined one. ROC AUC metric is chosen as the main one, and all model optimization is based on it. All metrics are calculated using the Scikit-learn package.

Chapter 4

Results & Discussion

4.1 Tox21 Classification Results

For each assay of Tox21, all five models were trained using 80%/10%/10% train / validation / test split. After that, the accuracy and AUC ROC metrics were calculated. Table 4.1 shows the AUC ROC results of each assay with different models on a test set. The underlined text in the table indicates which of the data representations has the highest ROC AUC score for each assay. Overall, models based on numerical data show better results for 9 out of 12 assays. Morgan Fingerprints, Mordred Descriptors, and RDKit Descriptors-based Random Forest classifiers have the highest AUC ROC in 4, 3, and 2 assays, respectively. The elEmBERT model is better in 2 assays, and the GIN model is better in 1 assay only.

The test AUC ROC of each model is used separately as a weight for every assay. Using this, the weighted ensemble results are calculated and shown in the Ensemble section of Table 4.2, along with the results of "Majority" and "Minority" votes. The bold text denotes ensemble algorithms that achieve the same or better results than the best-performing classification model. The table shows that a "Majority vote" improves the AUC ROC results in 9 out of 12 assays. On average, the score is elevated by 4.76%. The degree of improvement varies between 1% and 12.5% depending on the assay. In 5 assays, the weighted average achieves the same results as the best model of that assay but underperforms in the rest. "Minority vote" shows the worst results, Table 4.1: Tox21 AUC ROC Results of Each Data Representation on Test Set.

Assay	Models				
	RDKit	Morgan	Mordred	Graph	Elem

nr-ahr	0,8770	0,8770	0,8953	0,7294	0,7426
nr-aromatase	0,8661	0,8580	0,8528	0,7113	0,6644
nr-er	0,7350	0,7510	0,7576	0,8185	0,7718
sr-mmp	0,8640	0,8877	0,8524	0,7375	0,7930
sr-p53	0,9592	0,9592	0,9660	0,7137	0,7883
sr-atad5	0,8701	0,9401	0,8866	0,7755	0,7496
nr-ar-lbd	0,8961	0,9083	0,9328	0,8298	0,8249
nr-ar	0,7971	0,8212	0,8165	0,8263	0,8905
nr-ppar-gamma	0,9914	0,9080	0,8039	0,6151	0,9890
sr-are	0,8029	0,8112	0,8170	0,6572	0,8228
sr-hse	0,8134	0,8557	0,8021	0,7789	0,7174
nr-er-lbd	0,8635	0,9075	0,8491	0,8635	0,7627

with the AUC ROC score lower than the best-performing model for each assay. This result is expected, as the "Minority vote" is aimed to improve precision but reduces overall results.

The current state-of-the-art model TrimNet achieves an average ROC AUC of 0.860 [41]. In our results, the average ROC AUC is equal to 0.935, which outperforms it by 7.5%.

Table 4.2: Tox21 Ensemble AUC ROC Results of Each Data Representation on Test Set.

Assay	Majority	Minority	Weight
nr-ahr	0,9044	0,7675	0,8899
nr-aromatase	0,9171	0,7500	0,8661
nr-er	0,9441	0,8129	0,7989
sr-mmp	0,9066	0,8143	0,8824
sr-p53	0,9634	0,7967	0,9634
sr-atad5	0,9850	0,8333	0,9350
nr-ar-lbd	0,9159	0,8400	0,9002
nr-ar	0,9260	0,8559	0,8271
nr-ppar-gamma	0,9914	0,7969	0,9914
sr-are	0,9023	0,8192	0,8226
sr-hse	0,8898	0,8448	0,8557
nr-er-lbd	0,9825	0,8310	0,9075

4.2 Other Datasets Classification Results

Similarly to the Tox21, the classification performance was evaluated on five additional datasets. Table 4.3 shows ROC AUC score results on each dataset's test set. Because the SIDER dataset provides 27 classification tasks, its results show average values for all

metrics. Similarly to the Tox21 results, numerical data representations show the highest results in 3 out of 5 datasets. The best performance is achieved by the model, based on Mordred descriptors, followed by Morgan Fingerprints, RDKit Descriptors, and Graph models, respectively. The elEmBERT model shows a significant drop in performance compared to the Tox21 results.

The "Majority vote" and "Weighted vote" ensemble approaches increase the ROC AUC score in 2 out of 5 datasets. However, the improvement is less than 1%. Compared to TrimNet, our ClinTox and SIDER results are better by 0.6% and 0.2%. Even though other datasets do not show an improvement in ROC AUC, the "Majority vote" and "Weighted vote" show only a 1% difference from the best model results. In the case of ClinTox, the "Minority Vote" predicted only toxic labels, which made it impossible to calculate the ROC AUC score.

Table 4.3: Other datasets' AUC ROC results of each data representation and ensemble method on the test set.

Dataset	Data Representation					Ensemble		
	RDKit	Mordred	Morgan	Elem	Graph	Major	Minor	Weight
ClinTox	0,9542		0,4504	0,7070	0,7070	0,9542	-	0,9542
hERG B	0,7997	$\frac{0,9542}{0,9066}$	0,7706	0,4986	0,8211	0,8955	0,3538	0,8955
hERG K	0,8202	0,8389		0,4898	0,7559	0,8426	0,7295	0,8426
DILI	0,7313			0,4965	0,7917	0,7902	0,6966	0,7902
SIDER	0,5920	$\frac{0,7917}{0,5895}$	$\frac{0,8409}{0,7698}$ $\frac{0,6569}{0,6977}$	0,5392	0,5874	0,6376	0,5976	0,6376
Average	0,7795	0,8162		0,5462	0,7326	0,8240	0,5944	0,8240

The accuracy results are slightly better, as shown in Table 4.4. Similarly to previous results, numerical data representations show better results overall for all datasets. However, the "Majority vote" and "Weighted vote" improved the accuracy results in 3 out of 5 datasets. The degree of improvement varies from 0.001% in ClinTox to 2.1% in DILI.

Table 4.5 shows that all the ensemble methods improve the precision results in Table 4.4: Other datasets' accuracy results of each data representation and ensemble method on the test set.

Dataset	Data Representation					Ensemble		
	RDKit	Mordred	Morgan	Elem	Graph	Major	Minor	Weight
ClinTox	0,9091		0,8939	0,9015	0,9015	0,9091	0,9015	0,9091
hERG B	0,8182	$\frac{0,9091}{0,8939}$	0,8030	0,6212	0,8485	0,8788	0,6970	0,8788
hERG K	0,8201	0,8387		0,4900	0,7472	0,8424	0,6082	0,8424
DILI	0,7292			0,5000	0,7917	0,7917	0,6458	0,7917
SIDER	0,7385	$\frac{0,7917}{0,7453}$	$\frac{0,8409}{0,7708}$ $\frac{0,7569}{0,8131}$	0,7108	0,7099	0,7475	0,7260	0,7475
Average	0,8030	0,8357		0,6447	0,7998	0,8339	0,7157	0,8339

every dataset. The "Minority vote" performs especially well by achieving the same top precision in ClinTox and hERG Blockers, slightly improving the results of SIDER, and significantly improving the results of hERG Karim and DILI by 13% and 12.5%.

Table 4.5: Other datasets' precision results of each data representation and ensemble method on the test set.

Dataset	Data Representation					Ensemble		
	RDKit	Mordred	Morgan	Elem	Graph	Major	Minor	Weight
ClinTox	1,0000		0,9916	0,9832	0,9832	1,0000	1,0000	1,0000
hERG B	0,9362	$\frac{1,0000}{0,9787}$	0,9149	0,7872	0,9149	0,9787	0,9787	0,9787
hERG K	0,8098	0,8247	0,8380	0,5394	0,8395	0,8544	0,9718	0,8544
DILI	0,6923	0,7692	$\frac{0,8077}{0,7124}$	0,5385	0,7692	0,8077	0,9231	0,8077
SIDER	0,7564	$\frac{0,8006}{0,8746}$		0,7372	0,5938	0,7600	0,8075	0,7600
Average	0,8389		0,8529	0,7171	0,8201	0,8802	0,9362	0,8802

The ensemble techniques affect the recall metric the least, as illustrated in Table 4.6. Only SIDER results have shown a 2% improvement by the "Majority" and "Minority" approaches.

Table 4.6: Other datasets' recall results of each data representation and ensemble method on the test set.

Dataset	Data Representation					Ensemble		
	RDKit	Mordred	Morgan	Elem	Graph	Major	Minor	Weight

ClinTox	0,9084	0,9084	0,9008	0,9141	0,9141	0,9084	0,9015	0,9084
hERG B	0,8302		0,8269	0,7115	0,8776	0,8679	0,7077	0,8679
hERG K	0,8270		0,8430	0,4912	0,7089	0,8345	0,5628	0,8345
DILI	0,7826	$\frac{0,8846}{0,8486}$ $\frac{0,8333}{0,7255}$	0,7778	0,5385	0,8333	0,8077	0,6154	0,8077
SIDER	0,6900		0,7192	0,6791	0,6657	0,7435	0,6275	0,7435
Average	0,8076	0,8401	0,8135	0,6669	0,7999	0,8324	0,6830	0,8324

4.3 Combined Dataset Classification Results

According to Table 4.7, each data representation except elEmBERT achieves at least 78% ROC AUC and Accuracy on a combined dataset. This dataset used 26.5k molecules with labels from different toxicity types, showing good classification results. The Morgan Fingerprints-based Random Forest classifier achieves the best results for every metric. Ensembling the results improved Precision with the "Minority" approach and Recall with "Majority" and "Weighted" approaches.

Table 4.7: Combined dataset's metrics results of each data representation and ensemble method on the test set.

Dataset	Data Representation					Ensemble		
	RDKit	Mordred	Morgan	Elem	Graph	Major	Minor	Weight
Accuracy	0,8212	0,8291		0,5260	0,7899	0,8400	0,6653	0,8400
Precision	0,6937	0,7125		0,3617	0,6186	0,7065	0,9219	0,7065
Recall	0,8097	0,8156	$\frac{0,8464}{0,7451}$ $\frac{0,8341}{0,8434}$	0,3739	0,7845	0,8482	0,5353	0,8482
ROC AUC	0,8182	0,8257		0,4943	0,7884	0,8422	0,7244	0,8422

Chapter 5

Discussion

The initial tests on the Tox21 dataset provide valuable insights into different ML models and data representations and their efficiency in classifying molecules' toxicity. Numerical data-based models show higher ROC AUC scores in 9 out of 12 assays, meaning that they capture essential features of molecules. Despite using the same Random Forest architecture, different numerical data representations show the best results for specific assays. This outlines the difference in captured features, their importance for specific classification tasks, and their complementarity. Although less efficient in most assays, graph and token-based representation provide an important alternative to numerical data, using topological and morphological properties of molecules' SMILES representation.

Results of the ensemble approaches, especially the "Majority vote," further prove the importance of using several classification models. By treating the result of each classification model as a vote and selecting the most frequent one, the ROC AUC score was improved in 9 of 12 assays, averaging 4.76% per assay. The "Minority vote" shows lower results but improves the overall prediction safety and minimizes the chance of a false non-toxic prediction. The "Weighted vote" provides results that do not outperform individual classification models while not being focused on the safety of predictions. Using alternative strategies for weight calculation could potentially improve its results.

Metric analysis of other datasets shows similar results. Numerical data representations have the best ROC AUC score and accuracy results in every dataset. A model based on chemical descriptors from the Mordred package shows better performance compared to RDKit descriptors and Morgan fingerprints. The main reason could be the type of gathered data and features. Mordred calculates more than 700 two- and three-dimensional descriptors, while RDKit calculates only 43 two-dimensional ones. Morgan

fingerprints can also be considered as a two-dimensional feature, focusing on the similarity of fragments in a molecule. Although these data representations are complementary, Mordred descriptors may provide more important information, resulting in better results.

The results of ensemble methods on other datasets indicate their efficiency, but not to the extent observed in the Tox21 dataset. Out of 5 datasets, the ROC AUC and accuracy were improved in 2 and 3, respectively. However, the degree of improvement ranges from less than 1% to 2%, smaller than the average improvement in Tox21. Several reasons may cause this:

1. Class balance. Tox21 has a disproportionate class balance, with more than 90% of molecules being non-toxic. This may result in models like Random Forest favoriting the majority class and still having good results. It can be noticed that datasets improved by ensemble methods, such as ClinTox and hERG Blockers, also have a strong class imbalance. Decreased performance of individual models on balanced datasets could affect the efficacy of ensemble methods.

2. elEmBERT Performance. Apart from the SMILES representation of molecules, Tox21 provided atomic coordinates of the molecules, which were used to generate tokens for elEmBERT. For other datasets, however, the position of atoms had to be calculated from SMILES using Open Babel software. Because SMILES do not store that information, coordinates had to be approximated, which could affect the prediction results.

The metric that benefits the most from the ensemble techniques is precision. While all three methods positively affect the results of individual models, the "Minority vote" shows the biggest improvement. This result is expected, as the "Minority vote"

predicts a toxic label if at least one model predicted that. This results in "toxic" predictions appearing more often, increasing the number of True Positive predictions and decreasing the number of True and False Negative ones.

On the other hand, the recall was the least affected by the combination of the results. Only SIDER results have shown a 2% improvement by the "Majority" and "Minority" votes. The possible reason is that recall is more affected by the performance of the individual models, as it measures the ability of a classifier to identify all instances of the positive class.

The combined dataset, which includes more than 25k molecules and different types of toxicity, shows considerably good results. In contrast to other datasets, the Random Forest model based on Morgan Fingerprints achieves the best results in all metrics. One possible explanation is the difference in the most important descriptors across datasets. Morgan fingerprints, on the other hand, could provide similar information regardless of toxicity type. Despite having different toxicity types potentially caused by different physicochemical processes, the best-performing classifier achieves 84.64% accuracy and a ROC AUC score of 0.843. All ensemble results do not differ significantly from the results of the individual models. The only exception is precision, which is greatly improved by the "Minority vote." Overall, it can be assumed that it is possible to predict several toxicity types with a single classification model with considerably good accuracy.

Chapter 6

Conclusion

This project proposed and tested a novel drug toxicity prediction framework. It uses several representations of molecules' chemical structure, such as Morgan fingerprints, chemical descriptors from RDKit and Mordred libraries, graphs, and tokens. Corresponding Machine Learning classification models were built for each data representation: Random Forest, Graph Isomorphism Network, and Deep Bidirectional Transformer.

The predictions of all models were combined using three different methods. The "majority vote" selects the most frequent prediction, the "minority vote" checks if at least

one prediction result is "toxic", and the "Weighted vote" uses the results of each model on the validation set to weigh predictions.

The performance was evaluated on six different datasets, each providing a different toxicity type. Additionally, all six datasets were combined and evaluated as well. The ROC AUC score was chosen as the main evaluation metric, along with accuracy, precision, and recall being measured as well.

Results demonstrate that individual models based on the numerical data representations show better results, with models based on Morgan fingerprints and Mordred descriptors having the best overall performance. Depending on the dataset, the ROC AUC scores vary between 65.7% and 99.1%. The "majority vote" has the most impact on the ROC AUC results, increasing scores by at least 1%, all the way up to 12.5% in half of the datasets. The "minority vote" improves the precision results the most,

but severely underperforms in other metrics. The "weighted vote" usually performs better than the "minority vote" but does not outperform the "majority vote" or the best individual models in most of the cases.

Evaluation of the combined dataset shows that it is possible to gather a dataset with different toxicity types and achieve a ROC AUC score of at least 0.846. Ensemble techniques, however, had little to no effect on such a dataset.

Overall, the achieved results indicate that using different data representations can be beneficial to the classification results. Models, trained on several toxicity types can achieve tolerable classification results. Results can be improved further by changing the method of calculating atomic coordinates, which would improve the performance of the token-based model and the whole classifier as well. Additionally, a more accurate weight calculation algorithm can be developed to improve the "weighted vote" results.

Bibliography

[1] Jack W Scannell, Alex Blanckley, Helen Boldon, and Brian Warrington. Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery*, 11(3):191–200, 2012.

[2] Elina Petrova. *Innovation in the Pharmaceutical Industry: The Process of Drug Discovery and Development*, pages 19–81. Springer New York, New York, NY, 2014.

[3] Om Silakari and Pankaj Kumar Singh. *Concepts and experimental protocols of modelling and informatics in drug design*. Academic Press, 2020.

[4] F Peter Guengerich. Mechanisms of drug toxicity and relevance to pharmaceutical development. *Drug Metabolism and Pharmacokinetics*, 26(1):3–14, 2011.

[5] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6):463–477, 2019.

[6] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.

[7] D-S Cao, J-C Zhao, Y-N Yang, C-X Zhao, J Yan, S Liu, Q-N Hu, Q-S Xu, and Y-Z Liang. In silico toxicity prediction by support vector machine and smiles representation-based string kernel. *SAR and QSAR in Environmental Research*, 23(1-2):141–153, 2012.

[8] Maya Hirohara, Yutaka Saito, Yuki Koda, Kengo Sato, and Yasubumi Sakakibara. Convolutional neural network based on smiles representation of compounds for detecting chemical motif. *BMC Bioinformatics*, 19:83–94, 2018.

[9] Yuwei Miao, Hehuan Ma, and Junzhou Huang. Recent advances in toxicity prediction: Applications of deep graph learning. *Chemical Research in Toxicology*, 36(8):1206–1226, 2023. PMID: 37562046.

[10] Yuzhong Peng, Ziqiao Zhang, Qizhi Jiang, Jihong Guan, and Shuigeng Zhou. Top: Towards better toxicity prediction by deep molecular representation learning. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 318–325. IEEE, 2019.

[11] Jae Yong Ryu, Mi Young Lee, Jeong Hyun Lee, Byung Ho Lee, and KwangSeok Oh. Deephit: a deep learning framework for prediction of herg-induced cardiotoxicity. *Bioinformatics*, 36(10):3049–3055, 2020.

[12] Abdul Karim, Vahid Riahi, Avinash Mishra, MA Hakim Newton, Abdollah Dehzangi, Thomas Balle, and Abdul Sattar. Quantitative toxicity prediction via meta ensembling of multitask deep learning models. *ACS Omega*, 6(18):12306– 12317, 2021.

[13] Joelle Gola, Olga Obrezanova, Ed Champness, and Matthew Segall. Admet property prediction: the state of the art and current challenges. *QSAR & Combinatorial Science*, 25(12):1172–1180, 2006.

[14] Youjun Xu, Ziwei Dai, Fangjin Chen, Shuaishi Gao, Jianfeng Pei, and Luhua Lai. Deep learning for drug-induced liver injury. *Journal of Chemical Information and Modeling*, 55(10):2085–2093, 2015.

[15] Fang Du, Haibo Yu, Beiyan Zou, Joseph Babcock, Shunyou Long, and Min Li. hergcentral: a large database to store, retrieve, and analyze compoundhuman ether-a-go-go related gene channel interactions to facilitate cardiotoxicity assessment in drug development. *Assay and Drug Development Technologies*, 9(6):580–588, 2011.

[16] Hao Zhu, Todd M Martin, Lin Ye, Alexander Sedykh, Douglas M Young, and Alexander Tropsha. Quantitative structure- activity relationship modeling of rat acute toxicity by oral exposure. *Chemical Research in Toxicology*, 22(12):1913– 1921, 2009.

[17] Jiarui Chen, Yain-Whar Si, Chon-Wai Un, and Shirley WI Siu. Chemical toxicity prediction based on semi-supervised learning and graph convolutional neural network. *Journal of Cheminformatics*, 13(1):1–16, 2021.

[18] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.

[19] Jiarui Chen, Hong-Hin Cheong, and Shirley Weng In Siu. Bestox: A convolutional neural network regression model based on binary-encoded smiles for acute oral toxicity prediction of chemical compounds. In *Algorithms for Computational Biology: 7th International Conference, AlCoB 2020, Missoula, MT, USA, April 13–15, 2020, Proceedings*, pages 155–166. Springer, 2020.

[20] Tailong Lei, Youyong Li, Yunlong Song, Dan Li, Huiyong Sun, and Tingjun Hou. ADMET evaluation in drug discovery: 15. accurate prediction of rat oral acute toxicity using relevance vector machine and consensus modeling. *Journal of Cheminformatics*, 8:1–19, 2016.

[21] Minjun Chen, Vikrant Vijay, Qiang Shi, Zhichao Liu, Hong Fang, and Weida Tong. FDA-approved drug labeling for the study of drug-induced liver injury. *Drug Discovery Today*, 16(15):697–703, 2011.

[22] Nigel Greene, Lilia Fisk, Russell T Naven, Reine R Note, Mukesh L Patel, and Dennis J Pelletier. Developing structure- activity relationships for the prediction of hepatotoxicity. *Chemical Research in Toxicology*, 23(7):1215–1222, 2010.

[23] Jinghai J Xu, Peter V Henstock, Margaret C Dunn, Arthur R Smith, Jeffrey R Chabot, and David de Graaf. Cellular imaging predictions of clinical druginduced liver injury. *Toxicological Sciences*, 105(1):97–105, 2008.

[24] Vishan Kumar Gupta and Prashant Singh Rana. Ensemble Technique for Toxicity Prediction of Small Drug Molecules of the Antioxidant Response Element Signalling Pathway. *The Computer Journal*, 64(12):1861–1875, 02 2020.

[25] Kaitlyn M Gayvert, Neel S Madhukar, and Olivier Elemento. A data-driven approach to predicting successes and failures of clinical trials. *Cell Chemical Biology*, 23(10):1294–1301, 2016.

[26] Abdul Karim, Matthew Lee, Thomas Balle, and Abdul Sattar. Cardiotox net: a robust predictor for herg channel blockade based on deep learning meta-feature ensembles. *Journal of Cheminformatics*, 13:1–13, 2021.

[27] Shuangquan Wang, Huiyong Sun, Hui Liu, Dan Li, Youyong Li, and Tingjun Hou. ADMET evaluation in drug discovery. 16. predicting herg blockers by combining multiple pharmacophores and machine learning approaches. *Molecular Pharmaceutics*, 13(8):2855–2866, 2016.

[28] Monica Campillos, Michael Kuhn, Anne-Claude Gavin, Lars Juhl Jensen, and Peer Bork. Drug target identification using side-effect similarity. *Science*, 321(5886):263–266, 2008.

[29] Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548*, 2021.

[30] Leihong Wu, Ruili Huang, Igor V Tetko, Zhonghua Xia, Joshua Xu, and Weida Tong. Trade-off predictivity and explainability for machine-learning powered predictive toxicology: An in-depth investigation with tox21 data sets. *Chemical Research in Toxicology*, 34(2):541–549, 2021.

[31] Yue Wu, Zhichao Liu, Leihong Wu, Minjun Chen, and Weida Tong. Bert-based natural language processing of drug labeling documents: A case study for classifying drug-induced liver injury risk. *Frontiers in Artificial Intelligence*, 4:729834, 2021.

[32] Anusha Garlapati, Neeraj Malisetty, and Gayathri Narayanan. Classification of toxicity in comments using NLP and LSTM. In *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 16–21. IEEE, 2022.

[33] Greg Landrum. Rdkit: Open-source cheminformatics software. 2016.

[34] Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23(1-3):3–25, 1997.

[35] Mufei Li, Jinjing Zhou, Jiajing Hu, Wenxuan Fan, Yangkang Zhang, Yaxin Gu, and George Karypis. Dgl-lifesci: An open-source toolkit for deep learning on graphs in life science. *ACS Omega*, 6(41):27233–27238, 2021.

[36] Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3:1–14, 2011.

[37] Shokirbek Shermukhamedov, Dilorom Mamurjonova, and Michael Probst. Structure to property: Chemical element embeddings and a deep learning approach for accurate prediction of chemical properties. *arXiv preprint arXiv:2309.09355*, 2023.

[38] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[39] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S.

Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[40] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pretraining of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[41] Pengyong Li, Yuquan Li, Chang-Yu Hsieh, Shengyu Zhang, Xianggen Liu, Huanxiang Liu, Sen Song, and Xiaojun Yao. TrimNet: learning molecular representation from triplet messages for biomedicine. *Briefings in Bioinformatics*, 22(4):bbaa266, 11 2020.