Toxicity Prediction Based on Molecular Structure Using Machine Learning

Tristan William Tucker
University of Victoria
tristantw2005@gmail.com

Tejal Simran Cheema
University of Victoria
tejalcheema@gmail.com

Abstract—**Our project focuses on attempting to develop a machine learning model to predict the toxicity of molecules based on their molecular structure. In our testing we used two model archetypes, a Support Vector Machine (SVM) and a Neural Network, both trained using the Tox21 dataset [1], a catalog of over 10,000 molecules and their relative toxicities based on twelve distinct biological factors. Through our experiments, we found our best results were with using a Neural Network with CHEMBERT featurization of SMILES strings, LDA dimensionality reduction and SMOTEENN resampling. Our results show a positive correlation between molecular structure and toxicity, but found faults in attempting to build one general model to predict all the biological factors at once. To demonstrate an application of our models, we built an app allowing users to input the name of any molecule, and have the model output its predicted toxicity. Link to GitHub here.**

I. Introduction

Toxicology is an important field in medical science and often a large barrier for the synthesis of new materials and pharmaceuticals. It includes the study of pollutants, infectants, mutagens; disciplines of Oncology, Exposomics, and Toxicogenomics, as well as whole range of other fields regarding dangerous substances to humans. Our project aims to explore the complexities of this field by trying to build an AI model that can predict the toxicity of molecules through predicting specific events they would trigger in the body.

A. Motivation

The inspiration for starting into this project was a paper published on developing molecular structures using generative AI [2]. The AI model was trained on a series of Simplified Molecular Input Line Entry System (SMILES) strings, an efficient means of representing the structural geometry and atomic connectivity of molecules within a single string. [3] Their model output was new SMILES strings representing newly generated molecules. The main idea as to why one may want to do this is for material or pharmaceutical synthesis. The problem with this is that it is expensive to put together the resources necessary to synthesize all these different structures and test their viability or safety. If there were a way to predict the toxicity of certain molecules, we could eliminate a vast majority of harmful toxins and focus solely on the viable subset. Furthermore, a study from 2022 [4] by the American Society of Biochemistry and Moleculecular Biology claimed that 90% of drugs fail clinical trials; where reportedly

> ”around 30% were due to unmanageable toxicity or side effects.”

These failures cost pharmaceutical large sums of both time and money through testing, and research. With a model that could accurately predict these unwanted toxic effects, we can mitigate the amount of losses and allow for more efficient drug discovery.

Today, toxicology is expanding rapidly thanks to the new tools being developed in AI. Although the idea of using Machine Learning to predict the toxicity of molecules is not new. Certain implementations of AI in toxicity testing, such as In Silico Toxicology [5], has been gaining widespread adoption since the early 20th century. The fault with current methods however is largely due to the amount of data needed to train AI models to then predict toxicity, including information from quantum mechanics simulations, QSAR modeling, and knowledge of certain physicochemical properties of the molecules. In many situations, we do not have the ability to gather or simulate all of the required data of these molecules, especially considering the cases where generative AI is used to generate millions of hypothetical molecular structures to discover new drugs. Building a model that could predict the toxicity of a molecule using only its molecular structure would solve this problem.

B. Related Works

The dataset used in this project is the Tox21 dataset (Toxicology in the 21st centuary), whih catalogues of over 10,000 SMILES strings and their relative toxicity based on 12 tasks [1]. Originally introduced as a challenge dataset by the National Institute of Health (NIH) in 2014, Tox21 has since been widely used in machine learning-based toxicology research. One of the most notable models developed using this dataset is DeepTox, a deep learning model that won the Tox21 challenge [6].

The creators of DeepTox highlighted the potential of deep learning in toxicity predictions, emphasizing that neural networks are particularly well-suited for this task due to their ability to construct abstract chemical features [6]. This insight influenced our approach, motivating the use of neural networks as one of the models trained on the Tox21 dataset. Given the demonstrated success of deep learning within the Tox21 challenge, it gives validity to its usage in toxicology prediction.

Beyond deep learning approaches, other studies have explored alternative methods for improving toxicity classification on Tox21. One such study focused on the Structure-Activity Relationship (SAR) classification problem, a challenge stemming from the inherent class imbalance in toxicity datasets [7]. It used the Tox21 dataset and explored various resampling techniques to mitigate data imbalance. However, while our work compares Support Vector Machines (SVMs) and neural networks, their approach employed Random Forest as a base classifier, applying different resampling methods to improve model performance [7].

C. *Defining the Problem of Toxicity*

Dr. Stanley E. Manahan, a professor at the University of Missouri, wrote a book called Toxicological Chemistry, detailing the many ways in which toxicity has been defined over the years. This describes how toxicity had been commonly defined in terms of lethality or dosage; but why modern techniques aim to go further with the classification of toxic compounds by classifying them based ”according to the parts of the body affected or by toxic effect.” [8]. The following table below depicts the 12 distinct biological tasks included in the Tox21 dataset (i.e. what our model predicts):

TABLE I
Tasks of the Tox21 dataset.

Task Key	**Description**
NR-AR	Androgen Receptor: Protein that binds to androgens.
NR-AR-LBD	Androgen Receptor LBD: Binding of androgenic compounds.
NR-AhR	Aryl Hydrocarbon Receptor: Receptor in cell cytoplasm that detects Aryl Hydrocarbons.
NR-Aromatase	Aromatase: Enzyme that converts androgens to estrogens.
NR-ER	Estrogen Receptor: Protein that binds to estrogen.
NR-ER-LBD	Estrogen Receptor LBD: Binding of estrogenic compounds.
NR-PPAR-gamma	Peroxisome Protein that controls the regulation of fat storage and glucose.
SR-ARE	Antioxidant Response Element: Defense against oxidative stress.
SR-ATAD5	ATAD5 gene: Involved in DNA damage response - could help predict carcinogens and mutations.
SR-HSE	Heat Shock Element: Triggers stress-response proteins under heat or toxic stress
SR-MMP	Mitochondrial Membrane Potential: Used as an indicator for mitochondrial dysfunction.
SR-p53	p53 Protein: Tumor suppressor involved in DNA repair and cell growth regulation.

A majority of the items listed are either enzymes, proteins or receptors that serve the body in some function. Toxicity in this context is the malicious disruption of these specific systems. Toxicity is a hard problem to define as many chemicals can be harmful in different ways, and in fact, most medical drugs are listed as toxins for the reason that they may harm certain parts of the body despite being employed for good reasons [9]. Toxicity is thereby not a binary value, nor is it a continuous spectrum, so alternative systems of describing the toxicity of molecules are needed. For example, the U.S National Toxicology Program characterized toxicology into seven components [10]:

1) Cellular toxicology: Detrimental alteration of cells
2) Genetic toxicology: Alterations of DNA by toxicants
3) Carcinogenesis: Potential to cause cancer
4) Reproductive and developmental toxicology: Effects on reproductive organs and embryos
5) Renal toxicology: Effects on the kidneys
6) Pulmonary toxicology: Effects on the lungs
7) Immunotoxicology: Effects on the immune system

Many of the biological assays that are present in the Tox21 dataset are represented in one or more of these catagories. This is not to say that this is a perfect system, as there will be countless examples of chemicals and toxins that are extremely dangerous to humans that do not trigger any warnings and some common helpful endobiotic compounds that will. This is still one of the better ways of tracing toxicity, and the current focus for this paper.

II. Methodology

Predicting toxicity using the Tox21 dataset is a binary classification problem. While many binary classification tasks can be effectively addressed using traditional machine-learning models, the complexity of chemical structures presents additional challenges. Simpler models, such as Support Vector Machines (SVMs), may struggle to capture the intricate relationships between molecular features. Therefore, this project focuses on comparing SVMs and neural networks to determine which model type achieves superior performance.

The pipeline of our training process is illustrated in the flowchart below:

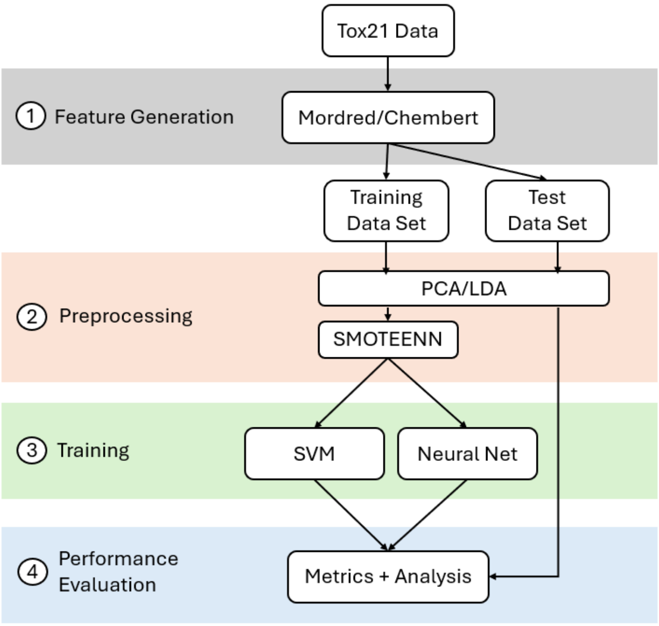

Fig. 1. Flowchart of model training.

The flowchart seen in figure I is split into four main sections. The basic process is explained below:

1) From the Tox21 dataset, we take the SMILES strings and featurize them into a vector embedding using either Mordred or CHEMBERT (*see Section on Dataset and Featurization*).
2) With the addition of vectorized molecules, we split the dataset into training and testing sets. For both sets we reduce the dimensionality using algorithms such as Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA). For the training set we use a technique called Synthetic Minority Oversampling Technique with Edited Nearest Neighbour (SMOTEEN) to increase visibility of the minority class.
3) After the preprocessing, we train our Support Vector Machines (SVMs) and neural network models.
4) Once models are trained, the test set is used to evaluate model performance based on the following key metrics: AUC, Precision, Recall and F1 Score.

A. Dataset and Featurization

As mentioned in the *Introduction*, the Tox21 dataset is a public dataset containing SMILES (Simplified Molecular Input Line Entry System) strings and 12 biological assay columns referred to as ”tasks”. A value of 1 in any task column indicates that the molecule triggered the assay, implying toxicity for that specific task.

There are two primary challenges with the Tox21 dataset. First, the dataset lacks predefined molecular features, as it only provides SMILES strings without additional structural or physicochemical descriptors. Second, the dataset exhibits class imbalance, favouring the value of 0 which indicates that most molecules do not trigger toxicity-related tasks.

To address the first challenge, we applied featurization, which is the process of transforming SMILES string into linear vectors or other favourable computer readable formats. We experimented with two main processes: Mordred [11] and CHEMBERT [12].

Mordred works by compiling an array of over 1000 integer values, each representing various structural features of the molecule. These features include, but are not limited to, the number of carbons, hydrogens, acid groups, basic groups, halogens, aromatic rings, and so forth. A complete list of the descriptors used can be found here.

CHEMBERT is an open source Bidirectional Encoder Representation Transformer (BERT) model designed to take SMILES strings and find the hidden relationships of the molecular structure. CHEMBERT provided far fewer features as Mordred (380 opposed to 1084) while also significantly improving model performance and training efficiency. Therefore, we decided to use CHEMBERT for all final models.

B. Preprocessing

To prepare the dataset for model training, we applied two key techniques: Dimensionality Reduction and Minority Class Resampling. Both techniques combined constructed a dataset with a suitable amount of features and enough instances of the minority class to train a decent model. The following sections detail these techniques individually.

1) Dimensionality Reduction:

Dimensionality reduction is a technique used to reduce the amount of features in large datasets while maintaining its most important information. Given that both Mordred and CHEMBERT produce a large quantity of features, not all of which are completely relevant for the different tasks, dimensionality reduction became an essential tool when training the models.

We experimented with two methods of dimensionality reduction algorithms: Principal Component Analysis (PCA) and Linear Dimensionality Reduction (LDA). These models were trained separately to explore their differences in results, and were compared to a baseline where no dimensionality reduction was applied.

PCA is a technique that identifies the most relevant principal components in the feature space, reducing dimensionality by discarding components that contribute the least variance. However, PCA is less interpretable, as it does not provide insight into which features are deemed more or less important. Given this, we decided to pivot towards using LDA, which is a supervised method that aims to find the projection that maximizes the separation (or discriminability) between different classes. Given its suitability for binary classification problems, LDA outperformed PCA and ultimately produced the best results.

2) Resampling:

Resampling is a method that attempts to fix imbalanced datasets by means of altering the exposure of certain classes within the dataset. There are two ways to go about this. One technique is oversampling which creates synthetic samples to increase the number of the minority class. The other is undersampling, which reduces the number of samples of the majority class.

Our models utilize a technique called SMOTEENN, a combination of the oversampling method SMOTE (Synthetic Minority Oversampling Technique) and the undersampling method ENN (Edited Nearest Neighbor). SMOTE generates synthetic samples by interpolating between nearby positive minority class instances. ENN removes samples that are misclassified by their k-nearest neighbors, primarily from the majority class, to improve class separability and reduce noise. This combination enhances model performance by balancing the dataset while filtering out ambiguous mislabeled samples.

Seeing the results in *Section III. Results*, models using SMOTEENN produced better results than without them.

C. Training

Following is a discussion of our training process, including insight into both our Neural Networks and Support Vector Machines.

1) Neural Networks:

The neural network models used in this project follows a simple design implemented using TensorFlow. The architecture consists of the following layers:

1) Input Layer: Receives the featurized molecular data derived from either Mordred or CHEMBERT representations of the molecules.
2) Normalization Layer: Applied to standardize input data, ensuring all features are scaled to a similar range.
3) First Hidden Layer: Contains 64 neurons and uses the ReLU (Rectified Linear Unit) activation function to introduce non-linearity and allow the model to learn more complex patterns
4) Second Hidden Layer: Contains 32 neurons, continuing to refine learned features and also uses ReLU activation function.
5) Output Layer: A single neuron with a sigmoid activation function to predict the probability of toxicity for each of the 12 tasks. The sigmoid function is used because the tasks are binary classification problems - a value of 1 represents toxicity, while 0 indicates non-toxicity.

To handle class imbalance inherent in the Tox21 dataset, we use Focal Loss as the loss function. Focal Loss mitigates the dominance of the majority class by down-weighting easy-to-classify samples, allowing the model to focus more on harder-to-classify instances. This makes it particularly effective for imbalanced datasets. For further details, refer to the Focal Loss paper [13].

Additionally, class weights are applied to further address the class imbalance by providing a higher weight to the minority class which reinforces the importance of making accurate predictions for the underrepresented class.

The class weights used in the training of both the SVM and Neural Networks were given by the following equation (Note: Each task had its own calculated set of class weights):

$$w_i = \frac{N}{2 \times n_i} \tag{1}$$

where:

- i is the class, either 1 or 0
- w_i is the class weight for class i,
- N is the total number of samples in the dataset,
- n_i is the number of samples belonging to class i.

2) SVMs:

The Support Vector Machines were implemented using sklearn's svm module which includes a variety of different kernel states and parameter options. From testing we found that the best results were gained by using the base SVM which employs a Radial Basis Function (RBF) kernel; good for complex, nonlinear data, which describes our input space for both Mordred and CHEMBERT.

D. Streamlit Integration

As an additional add-on to our project, a streamlit app was created to showcase a potential demo for a scenario where our models would be used by pharmaceutical companies. The models of all 12 tasks are saved and loaded into the app, where the user can select the task for a prediction.

The integration of the app is as follows:

1) User inputs the name of a molecule which exists in the PubChem database, a free chemical database containing over 100 million known compounds. [14]
2) Through the pubchem database we retrieve its representative SMILES string and parse it through CHEMBERT's API to generate features.
3) The features then undergo dimensionality reduction before being put through the models to generate predictions on toxic effect.
4) The predictions then get outputted for the users to view.

For further information regarding the app, refer to our Github page.

III. Results

In this section, we will discuss and compare our results using neural networks and SVMs for our binary classification problem.

For both neural networks and SVMs, different featurization, dimensionality reduction and resampling techniques were explored (*Refer to Section II. Methodology*). Ultimately, through experimentation, our best results for both model types had the following combination: CHEMBERT, LDA, and SMOTEENN.

The table below presents the averages scores of both model types across all 12 tasks. Although SVMs show a greater trend in Recall, the neural networks outperform in F1 score, Precicion, and AUC score. And since the F1 score represents the harmonic mean between Precision and Recall, it balances both metrics, making it particularly important in this context where minimizing false negatives is critical. Given this, the neural network models are deemed better for toxicity prediction.

TABLE II
Average metrics for models of each task

Model	AUC	Precision	Recall	F1
Neural Networks	87.01%	48.45%	48.88%	46.92%
SVMs	83.98%	24.93%	68.92%	35.24%

Figure 2 (below) showcases the precision-recall graph for all neural networks models of each task. It can be seen that each model varies in their results, where many seem to under perform significantly. The reason is due to the fact that we were trying to build a general neural network model to be used for each task. That is, the hyper-parameters were the same among all tasks. Based off the results, this proved to be a much more ambitious and challenging approach to the problem as each task exhibits different properties. Therefore, any future work should create neural network models unique to the task to produce optimal results.

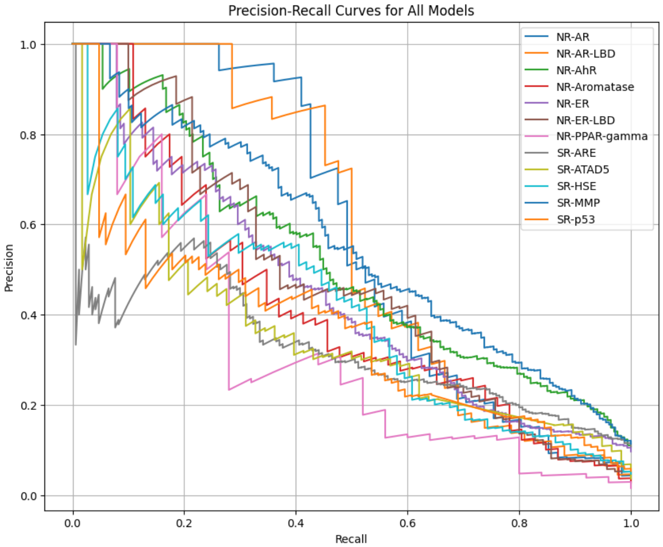

Fig. 2. Precision-Recall Graph for All Neural Net Models using LDA

The performance of our metrics found in *Table 1* for neural networks could be explained by the imbalanced of the dataset, combined with the fact that a general neural network model was utilized.

IV. Conclusion

This study compared the difference between SVMs and neural networks in building a general machine learning model that could predict the toxicity of molecules as defined in the Tox21 dataset. Our best results were gained through using a combination of CHEMBERT featurization, LDA dimensionality reduction and SMOTEEEN resampling for both the SVMs and Neural Networks.

Given the metrics we used, we claimed that the neural networks outperformed the SVMs for this given problem, but there is still great diffidence pertaining to the areas such as recall and precision. One reason for this could be our persistence in applying a general neural network architecture across all 12 tasks. As shown in *Figure 2*, this generalized approach proved ineffective, as each task was so unique. Future developments would disband this methodology and focus on building task specific models for each task. Therefore, until this additional work is completed, the conclusion of whether SVMs or neural networks are better suited for toxicity prediction across different tasks cannot be made.

Our models appeared to uphold to the long recognized trend of chemical Structure-Activity Relationship (SAR) [15]: which argues that the structure of a chemical can be predicted by its physical and chemical characteristics. We can make this claim by addressing the fact that by only using only a vectorized representation of a molecule's structure, we were able to receive an average AUC score of 85% between the two models. This implies that molecular structure alone carries enough information to make meaningful toxicity predictions.

References

[1] A. M. Richard, R. Huang, S. Waidyanatha, P. Shinn, B. J. Collins, I. Thillainadarajah, C. M. Grulke, A. J. Williams, R. R. Lougee, R. S. Judson *et al.*, "The tox21 10k compound library: collaborative chemistry advancing toxicology," *Chemical Research in Toxicology*, vol. 34, no. 2, pp. 189–216, 2020.

[2] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, "Automatic chemical design using a data-driven continuous representation of molecules," *ACS central science*, vol. 4, no. 2, pp. 268–276, 2018.

[3] E. J. Bjerrum, "Smiles enumeration as data augmentation for neural network modeling of molecules," *arXiv preprint arXiv:1703.07076*, 2017.

[4] D. Sun, "90% of drugs fail clinical trials," *ASBMB Today*, March 2022. [Online]. Available: https://www.asbmb.org/asbmb-today/opinions/031222/90-of-drugs-fail-clinical-trials

[5] A. B. Raies and V. B. Bajic, "In silico toxicology: computational methods for the prediction of chemical toxicity," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 6, no. 2, pp. 147–172, 2016.

[6] A. Mayr, G. Klambauer, T. Unterthiner, and S. Hochreiter, "Deeptox: toxicity prediction using deep learning," *Frontiers in Environmental Science*, vol. 3, p. 80, 2016.

[7] G. Idakwo, S. Thangapandian, J. Luttrell, Y. Li, N. Wang, Z. Zhou, H. Hong, B. Yang, C. Zhang, and P. Gong, "Structure-activity relationship-based chemical classification of highly imbalanced tox21 datasets," *J. Cheminform.*, vol. 12, no. 1, p. 66, Oct. 2020.

[8] S. E. Manahan, *Toxicological Chemistry and Biochemistry*, 1st ed. CRC Press, 2002. [Online]. Available: https://www.routledge.com/Toxicological-Chemistry-and-Biochemistry/Manahan/p/book/9781566706186

[9] S. Stevens, *Deadly Doses: A Writer's Guide to Poisons*. Cincinnati, OH: The Writer's Digest Books, 2004.

[10] N. T. P. (US), *National Toxicology Program: Annual Plan for Fiscal Year 1986*. National Toxicology Program, Public Health Service, Department of Health and . . . , 1987.

[11] H. Moriwaki, Y.-S. Tian, N. Kawashita, and T. Takagi, "Mordred: a molecular descriptor calculator," *Journal of cheminformatics*, vol. 10, no. 1, p. 4, 2018.

[12] H. Kim, J. Lee, S. Ahn, and J. R. Lee, "A merged molecular representation learning for molecular properties prediction with a web-based service," *Scientific Reports*, vol. 11, no. 1, p. 11028, 2021.

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[14] S. Kim, "Pubchem: A large-scale public chemical database for drug discovery sunghwan kim and evan e. bolton national center for biotechnology information, national library of medicine, national institutes of health, 8600 rockville pike, bethesda, md 20894, usa," *Open Access Databases and Datasets for Drug Discovery*, p. 41, 2023.

[15] B. T. Walton and T. Mill, "Structure-activity relationships in environmental toxicology and chemistry," pp. 403–404, 1988.