Scuola universitaria professionale
della Svizzera italiana

SUPSI

University of Applied Sciences and Arts of Southern Switzerland
Department of Innovative Technologies

Applied Case Studies of Machine Learning and Deep Learning in Key Areas 2

TOXICITY PREDICTION

Dyuman Bulloni
dyuman.bulloni@student.supsi.ch

Manuel Ippolito
manuel.ippolito@student.supsi.ch

Fabio Loddo
fabio.loddo@student.supsi.ch

21/05/2023

Table of Contents

List of Figures

1. Abstract

With drug safety being a critical aspect, there has been growing interest in chemical toxicology prediction within drug discovery research. In the past, researchers heavily relied on in vitro and in vivo experiments to assess the toxicity of chemical compounds. Nevertheless, these experiments are not only time-consuming and expensive, but they also raise ethical concerns due to the increased use of animal testing. Although traditional machine learning (ML) techniques have shown some promise in this field, the scarcity of annotated toxicity data remains a significant obstacle to enhancing model performance. We ought to expand the research and compare our results with the current state of the arts models.

2. Introduction

In this paper, we undertake a research task that involves an exploration of scientific papers available in the literature. Our objective is to identify 2-3 instances that share similarities with our own task and utilize the same dataset, serving as benchmark models. By leveraging the existing benchmarks, we aim to develop an engineered model capable of surpassing the performance scores achieved by these benchmarks, while utilizing the identical data splits employed in training the benchmark models.

The selection of benchmark models is a critical step in assessing the efficacy of our engineered model. By choosing instances that align with our task and data, we establish a reference point against which we can measure the advancements made by our model. Consequently, this comparison allows us to ascertain the effectiveness of our proposed approach in addressing the challenges posed by our specific task.

Our methodology involves a thorough examination of relevant scientific literature, where we identify and analyze various instances that pertain to our task domain. Through this extensive review, we seek to gain valuable insights into existing approaches, techniques, and methodologies employed by researchers in similar domains.

Furthermore, to ensure a fair evaluation of our model's performance, we adopt the data splits utilized by the benchmark models during their training phase. This approach allows for a direct comparison between the results obtained from our model and those achieved by the benchmark models, thereby establishing a meaningful evaluation framework. By surpassing the performance scores of the benchmark models, our research aims to contribute to the advancement of the field.

The papers we chose as reference are (Sharma, 2023) [1] and (Fang, 2022) [2].

3. Materials and Methods

3.1 Data

[3] The ClinTox dataset, introduced as part of this work, compares drugs approved by the FDA and drugs that have failed clinical trials for toxicity reasons. The dataset includes two classification tasks for 1491 drug compounds with known chemical structures: (1) clinical trial toxicity (or absence of toxicity) and (2) FDA approval status. List of FDA-approved drugs are compiled from the SWEETLEAD database, and list of drugs that failed clinical trials for toxicity reasons are compiled from the Aggregate Analysis of ClinicalTrials.gov (AACT) database.

3.2 Processing

We followed the pre-processing procedure we analyzed in the papers making the least possible variations to ensure comparable results. The Clintox dataset showed a very low number of null values.

Our initial cleaning process primarily involved the removal of non-valid rows, specifically those containing smiles strings that could not be converted into canonical smiles. By eliminating such rows, we ensured the consistency and integrity of the dataset, enabling subsequent analyses to be conducted on a reliable and valid subset of molecules. We then carefully examined the dataset for potential redundancies, seeking to eliminate duplicated molecules that could introduce bias or distort the results. Through this evaluation, we further refined the dataset, ensuring that each unique molecule was appropriately represented, thereby mitigating any inadvertent biases that may arise from redundant entries.

3.3 Modelling

3.3.1 Deep Purpose

One of the model we built for our pipeline is based on the Deep Purpose library [4]. We used the built-in deep neural network model on which we later performed hyper-parameter tuning. The library allow us to process the data and test our model on a series of different encodings. After some testing the best encodings that provided the better scores proved to be: CNN(Convolutional Neural Network on SMILES) and ESPF(Explainable Substructure Partition Fingerprint).

3.3.2 Convolutional Networks on Graphs

[5] Convolutional Networks on Graphs (ConvNets) have emerged as a powerful tool for learning representations and extracting meaningful features from graph-structured data. They provide a flexible and expressive framework for modeling complex relationships and interconnections, making them well-suited for various domains, including chemistry and drug discovery.

Molecular fingerprints encode structural and chemical information of molecules, capturing essential characteristics that influence their properties and activities. Traditional fingerprint generation methods, such as circular fingerprints, have been widely used but often lack the ability to capture fine-grained details and long-range interactions. ConvNets, on the other hand, offer a promising approach to overcome these limitations by leveraging the inherent graph structure of molecules.

The underlying principle of ConvNets on Graphs involves performing local operations on neighborhoods of nodes, allowing information to propagate through the graph. By employing learnable filters and convolutional layers, ConvNets can automatically learn relevant features and hierarchical representations directly from the molecular graph. This capability enables the extraction of discriminative and context-aware molecular fingerprints that capture both local and global structural patterns.

For this reason we choose to build a machine learning model based on this concept, borrowing the implementation from the DeepChem library [6].

3.4 Feature selection

The feaure selection metodologies were performed with the Deep Purpose deep neural network as base model since it was the most promising one.

3.4.1 Feature engineering

From our smiles strings we computed the following descriptors:

- Raw molecules
- LogP value
- Exact molecular weight
- Number rotatable bounds
- Heavy atoms
- Aromatic atoms
- AP

3.4.2 Recursive Feature Elimination (RFE)

In our model, we employ the Recursive Feature Elimination (RFE) technique, combined with a Random Forest model. RFE is a feature selection method that iteratively eliminates less important features from the dataset, based on their contribution to the predictive performance. By recursively removing features and assessing the impact on model performance, RFE helps identify the most informative features for the task at hand. The Random Forest model serves as the underlying classifier during the RFE process, allowing us to evaluate the importance of each feature based on its effect on the overall model performance.

3.4.3 Stratified k-fold Cross Validation with Shuffled Data

To evaluate the performance and robustness of our model, we used Stratified k-fold Cross Validation with shuffled data. This technique basically divides the dataset into k equally sized folds while preserving the class distribution proportions. Stratification ensures that each fold contains a representative sample of the different classes, which is particularly important in imbalanced datasets, as in our case.

We incorporated shuffling of the data before performing cross-validation. Shuffling the data helps eliminate any potential biases introduced by the original order of the dataset, ensuring that the performance evaluation is not affected by any inherent ordering or sequence. By shuffling the data before each cross-validation run, we enhance the randomness and reduce any dependencies on the original ordering of the samples. Finally, the Stratified k-fold Cross Validation with shuffled data allows us to obtain reliable and unbiased estimates of our model's performance.

4. Results and discussion

We evaluated the performance of our models using the following metrics: Area under Receiver Operating Characteristic curve (AUC-ROC), F1 score and balanced accuracy, precision, recall, accuracy.

The best model employed the two papers on ClinTox (clinical) were:

- A Deep Neural Networks trained with either Morgan fingerprint or with personalized SMILES Embeddings.
- A novel geometry-enhanced molecular representation learning method called GEM. The method consists of two main components: a geometry-based graph neural network (GeoGNN) architecture and various geometry-level self-supervised learning tasks.

The results obtained can be seen in Table 1.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC	Balanced Accuracy
DNN DeepPurpose with CNN SMILES embedding (ours)	0.971622 ± 0.00540541	0.84869 ± 0.0821908	0.857143 ± 0.0903508	0.850938 ± 0.0267221	0.999107 ± 0.0423747	0.920362 ± 0.0405656
DNN with SMILES embeddings	-	-	-	-	0.987 ± 0.019	0.955
ChemRL-GEM under random scaffold splitting	-	-	-	-	0.977 ± 0.019	-
Convolutional Graph Neural Network	0.495146	0.758915	0.241085	0.181172	0.627031	0.743818

Table 1: Machine Learning Model Results

The dashes reported in the table are due to the non-availability of these metrics in the reference papers. Also note that our confidence intervals take into account 2 standard deviations, such that about 95% of the point evaluation are considered, while the confidence intervals provided by the reference papers take into account just 1 standard deviation, in fact their standard deviations are close to half of the values reported by us. We preferred to report the values for 2 standard deviations as they are more representative of the real model performances.

Figure 1 and 2 shows the ROC-AUC curve and the Precision Recall curve over various runs of our model (DNN DeepPurpose with CNN SMILES embedding). For simplicity and readability reasons, just 5 runs are reported in the plot:

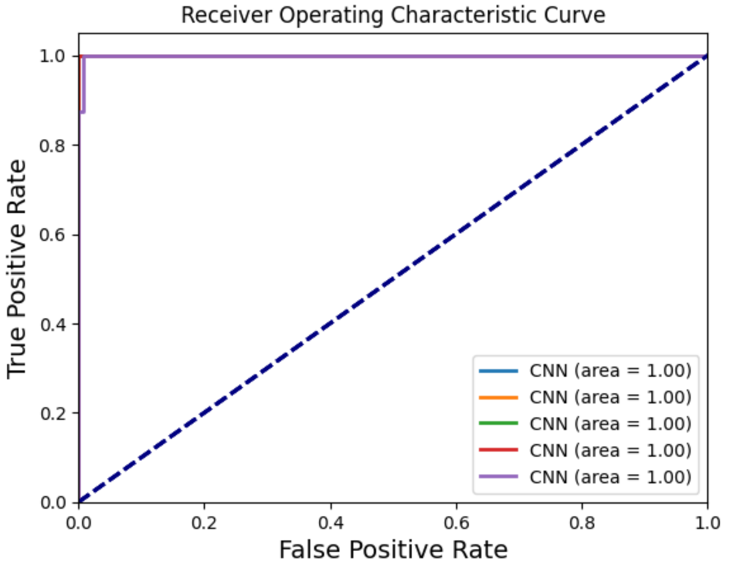

Figure 1: ROC-AUC curve over 5 runs of our model

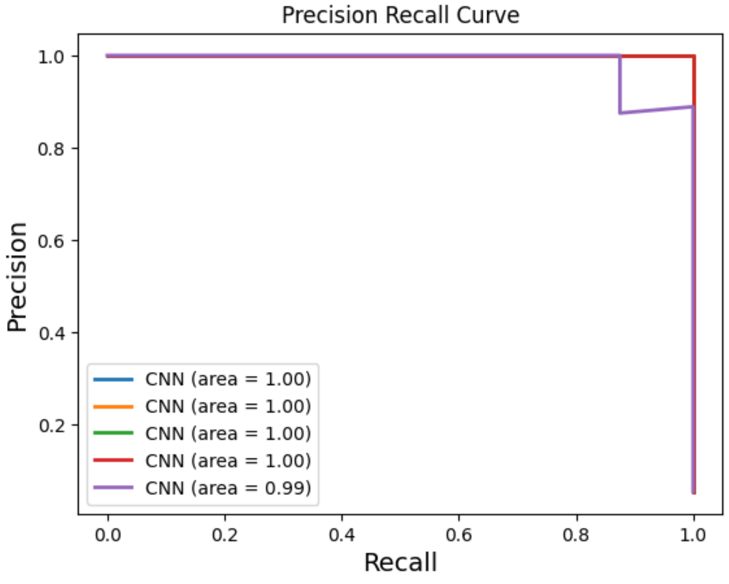

Figure 2: Precision-Recall curve over 5 runs of our model

Unbalanced datasets pose a pervasive challenge in the field of predictive toxicology. Irrespective of the specific platform employed, it is common to encounter unbalanced distributions of toxic and non-toxic examples. Among the datasets examined in this study, there exists a bias towards the ”non-toxic” class in ClinTox. Consequently, this imbalance exerts an influence on the AUC-ROC values, favoring a small fraction of accurate predictions for toxic or non-toxic outcomes. To address this issue, the balanced accuracy (BA) metric has been employed in most of the reference papers as a more appropriate measure for evaluating the performance of predictive toxicology models. By accounting for the dataset’s inherent imbalance, the BA metric offers a more representative evaluation of the model’s predictive capabilities in the domain of toxicology. We followed the indications given by the reference paper and used the same metrics and some other to have a more extensive view on the model performances.

The results we attained are comparable to the ones of the best models between our reference paper. Note that The first paper used also datasets about in vivo (RTECS) and in vitro (Tox21) toxicity other than the clinical (Clintox) data used by us, so it has probably better generalization performances. Same goes for the second paper, where the model was trained and test on a variety of different chemical tasks, not only toxicity classification.

5. Deployment

5.1 Prediction explanation

We selected a random molecule from the ones at our disposal, we then created a sample space.

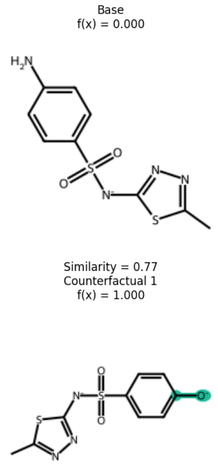

Figure 3: Check the molecule and a counterfactual graphically

The molecules seem to be similar in structure, but the counterfactual is toxic while the true molecule is not. The green higlighted atoms are the ones that made the prediction go towards classifying the molecule as toxic.

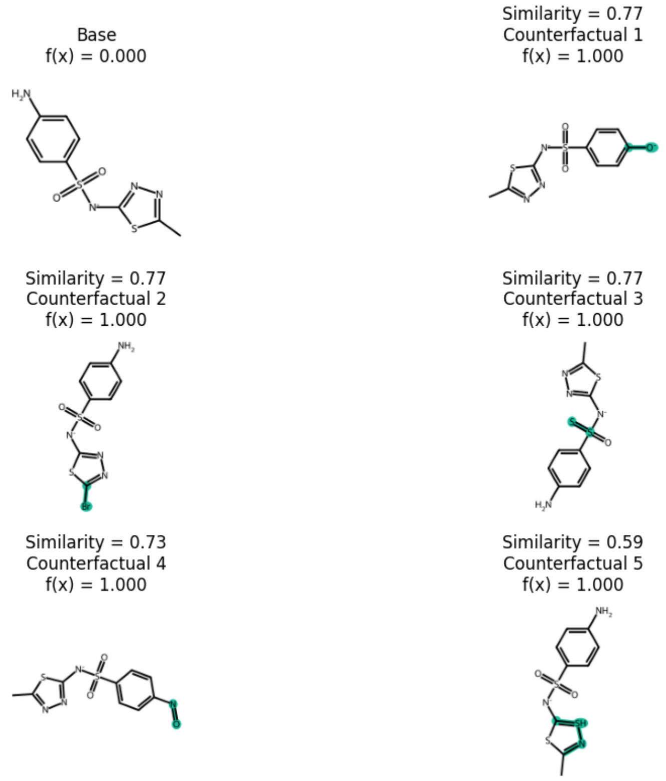

Figure 4: Let's take a look at more than one counterfactuals

The counterfactuals seem to have a similar structure with respect to the Base molecule. Again the green highlighted atoms are the ones that made the prediction go towards classifying the molecule as toxic. It seems that the atoms bond the outer rings are important for the predictions.

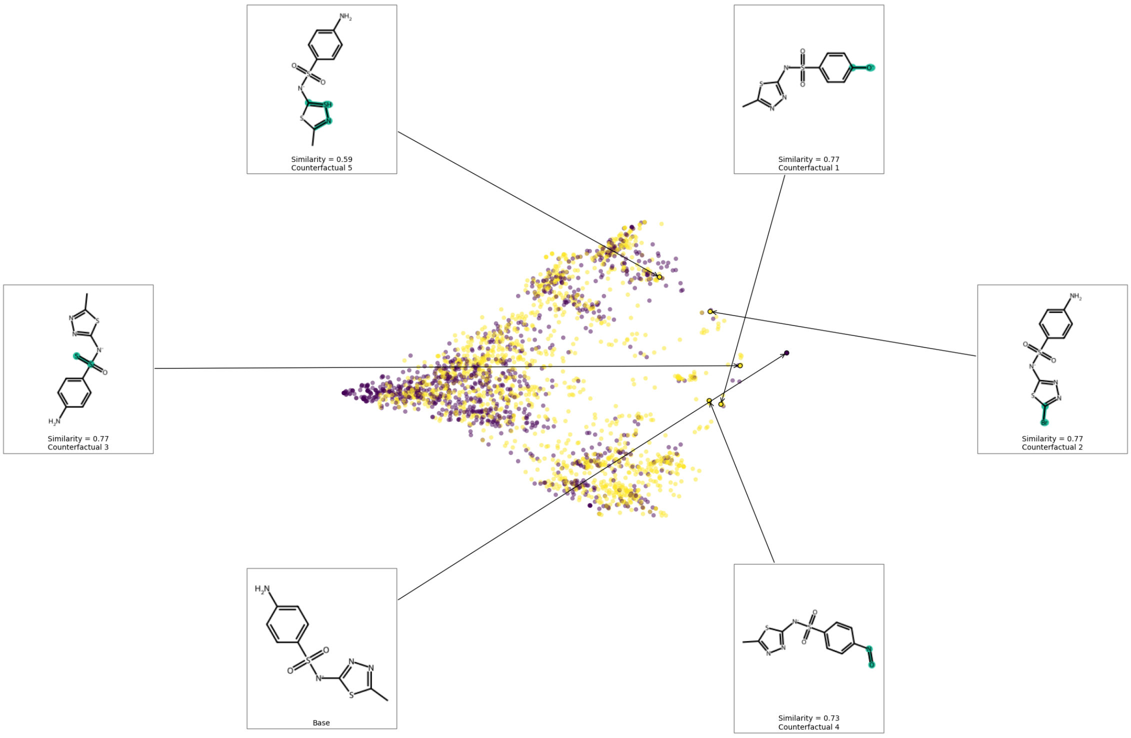

Figure 5: Chemical space around the example and annotations on given examples (PCA 2 components)

From the PCA 2 components representation, the base molecule and counterfactuals are quite near from each other, but the base molecule is outside of the counterfactuals sample space. The two classes are overlapped, but the chosen molecule seem to be a bit out of the sample space.

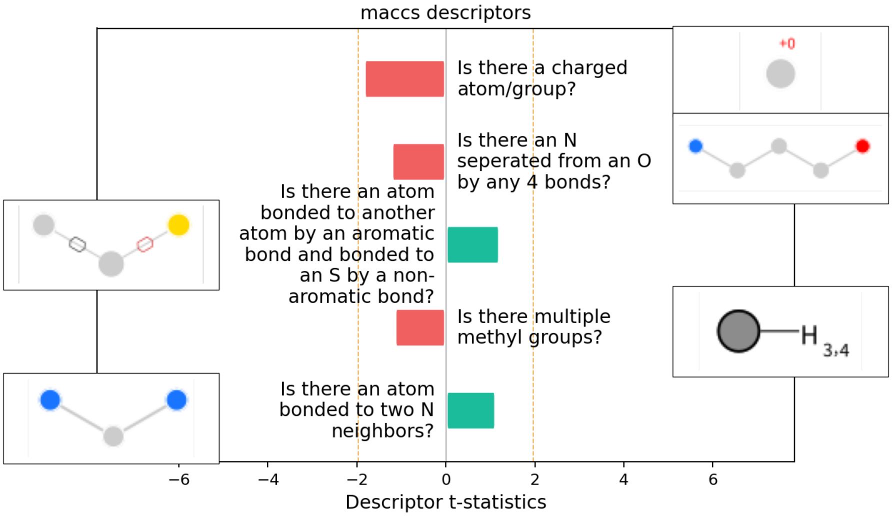

Figure 6: Explain using substructures

From the substructures explanation, It seems that to have a prediction=1 the most important substructure are: The presence of an atom bonded to another atom by abn aromatic bond and bonded to the S by a non-aromatic bonds. The presence of an atom bonded to two N neighbors.

The characteristics that are leading to a prediction=0 instead are:

- The presence of a charged atom/group.
- The presence of an N separated from an O by any 4 bonds.
- The presence of multiple methyl groups.

Looking at the local explanation above, between the Base molecule and Counterfactuals, the substructures indications doesn't give us information to think they are very different and belong to different classes, but again the one before was just a local explanation of 5 counterfactuals.

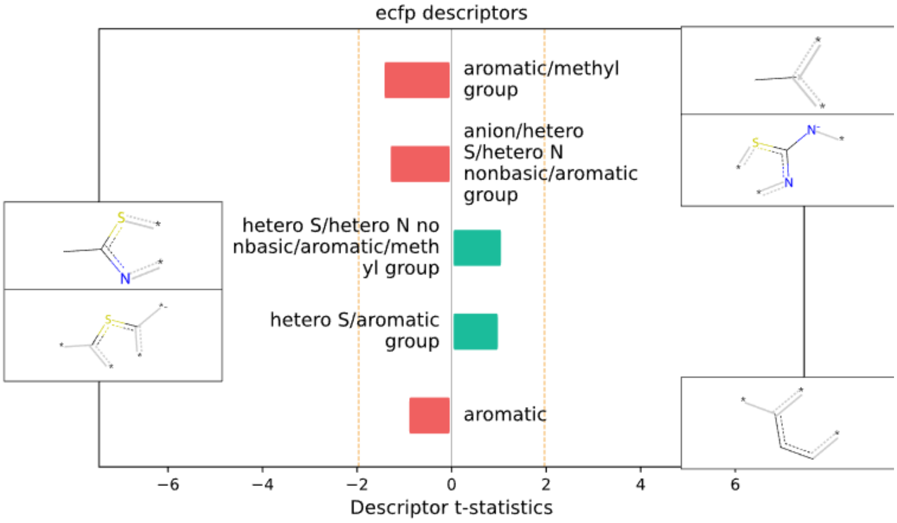

Figure 7: LIME explanation

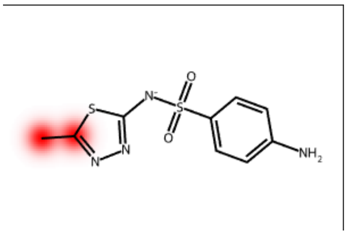

Figure 8

From the LIME explanation using ECFP as descriptor type, we see that there is not a single thing that can direct our prediction. The presence of a hetero S/hetero N nonbasic/aromatic/methyl group or an hetero S/aromatic group could lead to predicting class 1. The presence of an aromatic/methyl group, an anion/hetero S/hetero N nonbasic/aromatic group or an aromatic atom couls lead to predicting class 0.

The most important part seem to be the bonds with S N N ring highlighted in red in the image below. It is still not easy to determine by the structure of the molecule if it is going to be classified as 0 or 1.

5.2 Applicability Domain

Applicability domain was defined by uncertainty in predictive probabilities shown in the Results chapter.
The seed used is 124 and is common through all operations, meaning torch, torch.cuda for the

Neural Network training, numpy and random for minor processes, rdkit.utils.data_process for the fingerprint encoding and the kfold used for the cross-validation.
Dataset was split accordingly to [1], to guarantee comparison with the performances.
Following in Table 2, 3 and 4, the distribution of the chemicals for each set.

Table 2: Train Set Distribution

Feature	**Mean**	**Std**	**Min**	**Max**
Smiles Length	59.017	41.426	2.0	339.0
ExactMolWt	382.696	226.959	29.998	1881.071
NumRotableBonds	5.730	5.766	0.000	53.000
AromaticAtoms	7.984	6.588	0.000	40.000
HeavyAtoms	26.133	15.397	1.000	136.000
AP	0.318	0.237	0.000	0.920

Table 3: Validation Set Distribution

Feature	**Mean**	**Std**	**Min**	**Max**
Smiles Length	56.905	46.215	3.000	316.0
ExactMolWt	373.603	251.959	27.011	1753.637
NumRotableBonds	5.459	5.876	0.000	32.000
AromaticAtoms	8.291	6.483	0.000	30.000
HeavyAtoms	25.250	16.822	2.000	121.000
AP	0.346	0.231	0.000	0.882

Table 4: Test Set Distribution

Feature	**Mean**	**Std**	**Min**	**Max**
Smiles Length	62.378	43.545	9.000	263.0
ExactMolWt	392.884	226.226	45.993	1700.173
NumRotableBonds	5.736	5.475	0.000	37.000
AromaticAtoms	7.804	7.308	0.000	60.000
HeavyAtoms	27.236	16.123	2.000	122.000
AP	0.302	0.223	0.000	0.810

The sets are comparable in the distribution and applicability, which is aspected from the outstanding results for the prediction

6. Conclusions

We were able to build a model able to reach, and for some metrics also surpass, the state-of-the-art models available in literature for drugs toxicity prediction.
Please note that some of the reference models were trained on multiple datasets and tested on ClinTox dataset, so their model will probably be better at generalizing over unseen real-world data. Our proposed model is also easily replicable as it uses DeepPurpose as backbone, which is a well-known library for these applications, and it doesn't need much time to be trained and to customized. It would have been interesting to train our model on some other datasets besides ClinTox in order to obtain a more general view on its capabilities on various known benchmarks, but unfortunately these kind of data is often restricted in access and thus we opted to using this well-known dataset which is commonly used in literature.

References

[1] B. Sharma, V. Chenthamarakshan, A. Dhurandhar, *et al.*, "Accurate clinical toxicity prediction using multi-task deep neural nets and contrastive molecular explanations," *Scientific Reports*, vol. 13, no. 1, p. 4908, Mar. 2023, ISSN: 2045-2322. DOI: 10.1038/s41598-023-31169-8. [Online]. Available: https://doi.org/10.1038/s41598-023-31169-8.

[2] X. Fang, L. Liu, J. Lei, *et al.*, "Geometry-enhanced molecular representation learning for property prediction," *Nature Machine Intelligence*, vol. 4, no. 2, pp. 127–134, Feb. 2022. DOI: 10.1038/s42256-021-00438-4. [Online]. Available: https://doi.org/10.1038%2Fs42256-021-00438-4.

[3] Z. Wu, B. Ramsundar, E. N. Feinberg, *et al.*, "Moleculenet: A benchmark for molecular machine learning," *Chem. Sci.*, vol. 9, pp. 513–530, 2 2018. DOI: 10.1039/C7SC02664A. [Online]. Available: http://dx.doi.org/10.1039/C7SC02664A.

[4] K. Huang, T. Fu, L. M. Glass, M. Zitnik, C. Xiao, and J. Sun, "Deeppurpose: A deep learning library for drug-target interaction prediction," *Bioinformatics*, 2020.

[5] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, *et al.*, *Convolutional networks on graphs for learning molecular fingerprints*, 2015. arXiv: 1509.09292 [cs.LG].

[6] B. Ramsundar, P. Eastman, P. Walters, V. Pande, K. Leswing, and Z. Wu, *Deep Learning for the Life Sciences*. O'Reilly Media, 2019, https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837.