## **1. Mutagenicity literature review by Dr. Stankovic**

**Introduction**

Mutagenicity can be defined as “the permanent and transmissible changes in the amount or structure of the genetic materials of cells and organisms”. Thus, the mutagen is a physical or chemical agent which causes direct (or indirect) damage to DNA and consequently increases the frequency of mutations above the natural background level. Among the other methods, mutagenicity is most commonly estimated by the Ames method performed on *Salmonella typhimurium* bacterial strains [1]. Although from the perspective of *in vivo* experiments, the Ames test is the best choice, it can not be used for predicting a large number of compounds (since, for instance, it takes 2 days). Thus, the development of methods that can predict mutagenicity from the chemical structure has become the mainstream method nowadays. In other words, by using *in silico* models [2], the amount of Ames tests that are required to be carried out is reduced, leading to lowering of needed time and resources. These methods are of particular significance in a hit to lead optimization, as well as with gaining regulatory acceptance e.g. for the registration of industrial chemicals within REACH [3] or the assessment of impurities in pharmaceuticals [4]. Further, the importance of mutagenicity prediction is also indicated by the recently established Ames/QSAR International Challenge Project, where 12 QSAR vendors across the world have worked in collaboration to test and improve their Ames QSAR tools using a database of over 12000 molecules [5]. It is known that the predictive power of the positive Ames test is not greater than 90% [6]. Thus, due to the reproducibility of the Ames test, it is difficult to construct models with accuracy over 95%. However, it is still of great importance to strive for this accuracy.

The main idea of *in silico* prediction is that chemical structure (through the physical and chemical properties) is related to biological activity since it can affect biological interactions and therefore the toxicological mode of action. Although molecular representations are extremely diverse, the most widely used are molecular fingerprints and molecular descriptors. Molecular fingerprints are the way in which chemical structure is encoded (most commonly) as a vector of binary digits, whereas molecular descriptor is defined as the result of some procedure that transforms chemical information into a number. Further, during the years, many machine learning (ML) methods were performed in order to classify molecules as mutagenic or not, namely: partial least squares discriminant analysis (PLSDA), mixture discriminant analysis (MDA), support vector machines (SVM), K-nearest neighbors (kNN), C5.0, random forest (RF), stochastic gradient boosting (GBM), extreme gradient boosting (XGB), Naive Bayes (NB) classifier, shallow neural network (SNN), deep neural network (DNN), Gaussian Processes (GP), long short-term memory network(LSTM), logistic regression (LR).

One of the first investigations in this field was done by Helma et al [7]. They used C 4.5, rule learner PART, and SVM algorithms to estimate the mutagenicity of molecules from the carcinogenic potency database (CPDB). PART and SVM gave the best results, achieving accuracies of 0.78 in 10-fold CV. Moorthy et al [8] used RF paired with physicochemical and descriptors and MACCS fingerprints to investigate six different mutagenic and carcinogenic endpoints same database. From selected 52 descriptors, in the case of Ames mutagenicity, they compose model with MCC = 0.43 and accuracy of 0.72 in the test set.

` Votano et al [9] used SNN, DF, and kNN coupled with topological indexes to predict Ames mutagenicity from a few sources. All models showed similar area under the receiver operating characteristic curve (AUROC) 0.92±0.01 and accuracy 0.83±0.01 in external validation set, where SNN was slightly better than the other two.

Kazius, McGuire, and Bursi composed one of the widely-used dataset, known as Kazius or Bursi dataset[10]. Zheng et al [11] used SVM to predict mutagenicity of molecules from this datasets. The model showed an accuracy of 0.85 in external validation (CV). Zhang and Aires-de-Sousa [12] performed RF prediction of mutagenicity using 17 empirical physicochemical descriptors and up to 635 2D descriptors generated to describe bond properties. Involvement of the empirical descriptors slightly improved accuracy, i.e. from 0.84 to 0.85.

Hansen et al [13] composed another dataset of Ames mutagenicity. They used set of various molecular descriptors and SVM, GP, RF, and kNN algorithms. SVM showed the highest AUROC in 5-fold CV, 0.86. Seal et al [14] used NB, SVM (precisely SMO), DT (precisely C 4.5), and RF to predict mutagenicity of the Bursi mutagenicity dataset and the Hansen dataset. The metrics of the model increase as the dataset is larger. They found that RF performed best. Accuracy and AUROC for the overall dataset were 0.795 and 0.892, respectively. They also predict the mutagenicity of molecules from CPDB, as the external validation set. Values for metrics were 0.901 and 0.968. Warning: If somebody is going to read the paper...They also did predictions on another external set. They obtained 0.640 and 0.646. That is mixing apples with oranges and it is not valid. But, what concerns me is if they dropped duplicates from the external set since 0.901 and 0.968 are too high. We can check. They also predict mutagenicity of 1410 approved drugs and 66 withdrawn drugs from the DrugBank database, as well as 22080 metabolite data which were taken from the recently published ZINC Data sets. We can also do that, but problem is that some drugs can be slightly mutagenic. Banerjee et al [15] used RF to predict the mutagenicity of chemicals from MACCS and Morgan fingerprints. The accuracy was 0.84 on cross-validation (Hansen dataset) and 85.00% on external (CPDB dataset) validation. Further, the AUC–ROC scores of cross-validation and external validation were 0.90 and 0.9, respectively. Zhang et al [16] developed a gradient boosting algorithm and use it with ECFP fingerprints to predict toxicity within Tox21 datasets and Hansen mutagenicity dataset. They obtained averaged AUROC of 0.836 in CV and 0.786 in external validation. They stated their GBM algorithm performed better than XGB, SVM, DNN, and RF from the point of view of averaged metrics values. Karim et al [17] used a hybrid model composed of SNN and DT with 2D descriptors to predict the mutagenicity of the Hansen dataset. AUCROC was 0.879. Li et al [18] used a graph convolution neural network to classify mutagenicity. They used the Hansen dataset for training and constructed external validation set from 3 databases. Several models were constructed and the best showed up to 0.810 of accuracy and 0.878 of AUROC in 10-fold CV and 0.766 of accuracy and 0.8382 of AUROC in external validation. Also, they combined SVM, RF, kNN, and XGB with fingerprints obtained from their model. The best classification with AUROC of up to 0.953 and 0.835 and accuracy of 0.887 and 0.766 in 5-fold CV and external validation, respectively were obtained with SVM. However, in the same paper, they showed that SVM with ECFP fingerprints, albeit having lower values in CV (0.876 and 0.818), has better performance in the external set (0.866 and 0.799).

Xu et al[19] used five models, SVM, C 4.5, SNN, kNN, and NB, and five classes of fingerprints, i.e. CDK (also known as FP), Estate, MACCS, PubChem, and SubFP fingerprints to predict mutagenicity from 5 different sources. They found that the best results are obtained by SVM and kNN combined with MACCS and PubChem fingerprints (accuracy 0.927-0.980, AUROC 0.924-0.970). Also, they stated that models which used fingerprints are comparable with those using molecular descriptors. Chu et al [20] used several combinations of 13 molecular properties (MP) and ECFP and FCFP figerprints, i.e. ECFP, FCFP, MP, ECFP+FCFP, MP+ECFP, MP+FCFP, MP+ECFP+FCFP, and several algorithms, i.e. PLSDA, MDA, SVM, kNN, C5.0, RF, GBM, XGB to predict mutagenicity of Xu’s Ames dataset. They found that SVM, RF, and XGB show better results than other algorithms and that although MP cannot be used alone to predicts mutagenicity, their show significant importance in models with fingerprints. RF had best accuracy (0.78) and AUCROC (0.84) in external validation set. They also showed that, out of the 36 curated Class A mutagenic chemical which were correctly predicted by one or fewer models within the Ames/QSAR International Challenge Project, their models can correctly predict the mutagenicity of 28.

Chakravarti and Alla [21] compose the largest know dataset of Ames mutagenicity and developed a model for predicting Ames mutagenicity based on LSTM and two different molecular linear notation, simplified molecular-input line-entry system (SMILES) and molecular linear notation by circular traverse (MLNCT). Also, they predict mutagenicity using ECFP fingerprints combined with DNN and LR. Models showed similar values of 10-fold CV prediction metrics (accuracy 0.858-0.867, AUROC 0.931-0.936), as well as for those for external set (accuracy 0.871-0.890, AUROC 0.938-0.942).

Helma et al [22] compose dataset from the Bursi dataset, Hansen dataset, and EFSA dataset [23]. Further, they used kNN (more precisely lazy structure-activity relationships algorithm), RF, LR, DNN, SVM with MolPrint2D fingerprints calculated with the OpenBabel cheminformatics library and 1D and 2D CDK descriptors calculated with the PaDEL-Descriptors program. Metrics for predictions with fingerprints and descriptors were comparable. The best performance showed DNN (accuracy 0.85) and RF (accuracy 0.84), both with CDK descriptors.

Polarizability (*α*) is the linear response of a matter to an external electric field, in which an electric dipole moment is induced. In general, polarizability is a three-dimensional tensor. However, in the majority of considerations, the average of tensor’s diagonal elements is used (<*α>*= 1*=*3 *×* (*α*xx+*α*yy+ *α*zz). Moreover, polarizability is an important property, since it defines the strength of dipole-induce dipole and dispersive interactions.

Tandon et al [24] stated that contemporary computational research overlooks the importance of twentieth-century findings of electronic interactions between a receptor and a ligand. More precisely, most of the current methods for estimation of biological activity do not take into consideration of polarizability. Although considered *α* as a sum of atomic polarizabilities, they showed that it can be used for the prediction of various types of biological activities, where *R*2 can be even 0.972. However, this approach has two drawbacks, and that is that neither different isomers, nether neutral species form ions can be distinguished. On the other hand, it is well-known that mutagenic activities of isomers of the different molecule can be in interval of a few orders of magnitudes [25]. Since dipole-induce dipole and dispersive interactions can govern the strength of interactions between mutagenic molecules and enzyme that activate it, good correlation between mutagenicity of isomers of some molecule and polarizabilities which computed by quantum chemistry (QC) methods can be found [26-28].

Molecular properties calculated using QC methods are rarely used as descriptors, due to the cost of their calculations. However, in recent years there were several attempts to develop ML procedures by which these quantities can be calculated for a reasonable amount of time and without (significant) loss of accuracy. This opens the possibility to estimate properties using ML methods and then use these results to predict toxicity by another ML algorithm.

Moreover, the prediction of polarizability as a molecular property demands an even larger computation cost than the majority of other properties. Precisely, high precision methods scale with six power of system size and the accurate calculation can be done only for molecules with about a dozen of atoms. Wilkins et al [29] developed a method based on symmetry-adapted Gaussian process regression (SA-GPR) and the smooth overlap of atomic positions (λ-SOAP) descriptors for the prediction of polarizability. They train it on the QM7b database [] which contains 7211 molecules with up to 7 “heavy” atoms (C, N, O, S, Cl) which polarizabilities are calculated at CCSD/aug-cc-pVTZ level of theory. They demonstrated that the model has better accuracy than QC methods with moderate precision. Also, they showed that their methods scale well with an increase in the molecule size. This is important since high precision calculations, in our cases, can easily require several thousands of CPU hours and up to a few TB of RAM.

The aim of this paper is to examine to which extent molecular properties calculated from polarizability tensor can improve the prediction of Ames mutagenicity. The method of Wilkins et al will be used to predict polarizability tensor. The values of *α*xx, *α*xy, *α*avg, and Δ*α* will be calculated and used with other molecular descriptors to predict Ames mutagenicity.

**Materials and methods**

*Data Preparation*

In this paper, this and that dataset was/were used. In order to obtain reliable results, inorganic molecules (organometallic and mixtures) were removed. After that, molecules with unspecified stereochemistry were removed and the molecules were standardized with the InChI key. Finally, the salt fragments were removed and any duplicates were identified and removed based on the InChI key.

Xu 2012 Compounds with molecular mass less than 40 or more than 800 are omitted from the dataset.

Li 2021 To build predictive models more robustly, we filtered mixtures, polymers, compounds containing heavy metals, and compounds with less than 3 carbon atoms.

***Data compilation***

***How we split the data?***

Chu et al Dataset was split into training and test sets using a few steps. First, the dataset was clustered into 1000 clusters using extended connectivity fingerprints (ECFP, diameter = 6). Then, where it is possible, a mutagenic and a non-mutagenic representative closest to the center of the cluster were selected for the test set. Molecules from clusters with only one representative were put into the test set. Mutagens from the clusters with two representatives were put into the test set until the threshold of 500 molecules. The non-mutagens from the rest of the clusters were put into the test set. Finally, any molecules not in the test set were put into the training set. (Not sure, what they have done. Result is test set of 7617 molecules and training of 731 (not balanced subset) + 234 (balanced subset))

Using the training set of each predictor data set, an applicability domain was defined by tracking the property range and analyzing the optimum prediction space. The molecules from the test set were then filtered for model applicability by analyzing them against the defined applicability domain. Any molecules which did not pass the model applicability filter were swapped for the next molecule closest to the cluster center within the same cluster with the same mutagenicity label from the training set. The applicability domain was then defined again via the same process until no more swaps could be made. Any molecules which did not pass the model applicability filter were then removed from the test set and placed back into the training set.

In order to find the optimized AD, four different thresholds (ranging from 0.17 to 0.14 with step size of 0.01) were determined according to the average Tanimoto coefficient of the training set (0.18)

AlogP vs Molecular weight t-SNE

***Which descriptor?***

The molecular descriptors were scaled to a range between 0 and 1.

Quantile transformation to a uniform distribution (Helma 2021)

***Data pre-processing***

*How we have chosen descriptors?* *For instance…* Predictors with variance lower than 0.1 were removed. Furthermore, descriptors with pair-wise absolute correlations over 0.9 were identified. For each pair, the descriptor with a higher average correlation with the rest of the set was removed.

Zhong et al (dx.doi.org/10.1021/tx4000182)

Pearson correlation analysis and stepwise regression method based on the whole data set were used to find the best combination of descriptors for modeling.

Chakravarti and Alla L1 regularization/Lasso regression

Karim et al gini importances (if I am correct)

***Algorithm selection***

*Model building and performance assessment*

5- or 10-fold cross-validation on the training set for each of the predictor data sets???

Zhang et al Nested cross validation. Bayesian optimization for global optimization of the ML algorithm hyperparameters.

The performance of classification models was assessed by the following metrics:

Sensitivity = (TP) / True

Specificity = (TN) / False

Accuracy= (TP + TN) / Total

balanced accuracy= (Sensitivity + Specificity) / 2

Area under the Receiver Operating Characteristic curve (AUROC)

Kappa= TP + TN) - (True × Positive + False × Negative)) / (1 + (True ×Positive + False × Negative))

Matthews correlation coefficient (MCC) Text

Description automatically generated

Shape

Description automatically generated (only in 1 paper)

*Model validation via y-randomization and model robustness*

Model validation via y-randomization was performed by random shuffling of the column with mutagenicity before training for 3 (5, 10) times for each of the selected models. Model robustness was estimated by Z score:

Z = (Kappa original training – Average(Kappa y-randomized training)) / SD(Kappa y-randomized training)

When the original model is valid, the performance of the y-randomized model is greatly reduced in comparison, which can be observed by a high Z score (Z *>* 3).

*OECD QSAR guidelines*

In this paper, we followed recommendation from the OECD QSAR Guidelines for structure-activity modelling (Table …)

http://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm

Example form one of the paper

Table

Description automatically generated

**References**

[1] Ames B: The detection of environmental mutagens and potential. Cancer 1984, 53:2030–2040.

[2] C.Xu,F.Cheng,L.Chen,Z.Du,W.Li,G.Liu,P.W.Lee,Y.Tang,In silico prediction of chemical ames mutagenicity, J. Chem. Inf. Model. 52 (2012) 2840–2847.

[3] European Chemical Agency (2017). Guidance on Information Requirements and Chemical Safety Assessment. Chapter R.7a: Endpoint Specific Guidance. European chemical agency guidance on information requirements. Helsinki: ECHA. doi:10.2823/337352.

[4] ICH (2017). ICH guideline M7(R1) on Assessment and Control of DNA Reactive (Mutagenic) Impurities in Pharmaceuticals to Limit Potential Carcinogenic Risk. London: European Medicines Agency, Committee for Human Medicinal Products. EMA/CHMP/ICH/83812/2015.

[5] M. Honma, A. Kitazawa, A. Cayley, R.V. Williams, C. Barber, T. Hanser, R. Saiakhov, S. Chakravarti, G.J. Myatt, K.P. Cross, et al., Improvement of quantitative structure-activity relationship (qsar) tools for predicting ames mutagenicity: outcomes of the ames/qsar international challenge project, Mutagenesis 34 (2019) 3–16.

[6] Mortelmans K, Zeiger E: The ames salmonella/microsome mutagenicity assay. Mutat Res 2000, 455(1–2):29–60.

[7] [Helma 2004] C. Helma, T. Cramer, S. Kramer, L. De Raedt, Data Mining and Machine Learning Techniques for the Identification of Mutagenicity Inducing Substructures and Structure Activity Relationships of Noncongeneric Compounds, *J. Chem. Inf. Comput. Sci.* **2004,** *44,* 1402-1411.

[8] N.S Hari Narayana Moorthy, Surendra Kumar, Vasanthanathan Poongavanam, Classification of carcinogenic and mutagenic properties using machine learning method, Computational Toxicology 3 (2017) 33–43. <http://dx.doi.org/10.1016/j.comtox.2017.07.002>.

[9] [Votano 2004] J. R. Votano, M. Parham, L. H. Hall, L. B. Kier, S. Oloff, A. Tropsha, Q. Xie, W. Tong, Three new consensus QSAR models for the prediction of Ames genotoxicity, Mutagenesis vol. 19 no. 5 pp. 365--377, 2004, doi:10.1093/mutage/geh043.

[10] Kazius, J.; McGuire, R.; Bursi, R. Derivation and Validation of Toxicophores for Mutagenicity Prediction. *J. Med. Chem.* **2005**, *48*, 312-320

[11] M. Zheng, Z. Liu, C. Xue, W. Yhu, K. Chen, X. Luo, H. Jiang, Mutagenic probability estimation of chemical compounds by a novel molecular electrophilicity vector and support vector machine, Bioinformatics, Vol. 22 no. 17 2006, pages 2099–2106.

[12] Q.-Y. Zhang, J. Aires-de-Sousa, Random Forest Prediction of Mutagenicity from Empirical Physicochemical Descriptors, *J. Chem. Inf. Model.* **2007,** *47,* 1-8.

[13] K. Hansen, S. Mika, T. Schroter, A. Sutter, A. tar Laak, T. Steger-Hartmann, N. Haienrich, K-R. Muller, Benchmark Data Set for in Silico Prediction of Ames Mutagenicity, *J. Chem. Inf. Model.* **2009,** *49,* 2077–2081.

[14] Abhik Seal, Anurag Passi, UC Abdul Jaleel, Open Source Drug Discovery Consortium and David J Wild, In-silico predictive mutagenicity model generation using supervised learning approaches, Journal of Cheminformatics 2012, 4:10.

[15] Banerjee P, Eckert AO, Schrey AK, Robert P (2018) ProTox-II: a webserver for the prediction of toxicity of chemicals. Nucleic Acids Res 46(W1):W257–W263. <https://doi.org/10.1093/nar/gky318>

[16] Zhang J, Mucs D, Norinder U, Svensson F (2019) LightGBM: an effective and scalable algorithm for prediction of chemical toxicity-application to the Tox21 and mutagenicity datasets. J Chem Inf Model 59(10):4150–4158. https://doi.org/10.1021/acs.jcim.9b00633

[17] A. Karim, A. Mishra, M. A. Hakim Newton, A. Sattar, Efficient Toxicity Prediction via Simple Features Using Shallow Neural Networks and Decision Trees, ACS Omega 2019, 4, 1874-1888. 10.1021/acsomega.8b03173

[18] S. Li, L. Zhang, H. Feng, J. Meng, D. Xie, L. Yi, I. T. Arkin, H. Liu, MutagenPred‑GCNNs: A Graph Convolutional Neural Network‑Based Classification Model for Mutagenicity Prediction with Data‑Driven Molecular Fingerprints, [Interdisciplinary Sciences: Computational Life Sciences](https://link.springer.com/journal/12539) volume 13, pages 25–33 (2021). <https://doi.org/10.1007/s12539-020-00407-2>

[19] C. Xu, F. Cheng, L. Chen, Z. Du, W. Li, G. Liu, P. W. Lee, Y. Tang, In silico Prediction of Chemical Ames Mutagenicity, J. Chem. Inf. Model. 2012, 52, 2840-2847. dx.doi.org/10.1021/ci300400a

[20] Charmaine S.M. Chu , Jack D. Simpson, Paul M. O’Neill, Neil G. Berry, Journal of Molecular Graphics and Modelling 109 (2021) 108011, <https://doi.org/10.1016/j.jmgm.2021.108011>

[21] S. K. Chakravarti, S. R. M. Alla, Descriptor Free QSAR Modeling Using Deep Learning with Long Short-Term Memory Neural Networks, *Front. Artif. Intell.* 2:17. doi:10.3389/frai.2019.00017

[22] C. Helma, V. Schöning, J. Drewe, P. Boss, A Comparison of Nine Machine Learning Mutagenicity Models and Their Application for Predicting Pyrrolizidine Alkaloids, Frontiers in Pharmacology Volume 12 , Article 708050. doi: 10.3389/fphar.2021.708050

[23] “Guidance on the Establishment of the Residue Definition for Dietary Assessment: EFSA Panel on Plant Protect Products and Their Residues (PPR).” EFSA J., 14. 1–12. doi:10.2903/j.efsa.2016.4549

[24] Tandon, H.; Ranjan, P.; Chakraborty, T.; Suhag, V. Polarizability: a promising descriptor to study chemical–biological interactions. *Molecular diversity* **2021**, *25*, 249–262.

[25] Takamura-Enya, T.; Suzuki, H.; Hisamatsu, Y. Mutagenic activities and physicochemical  
properties of selected nitrobenzanthrones. *Mutagenesis* **2006**, *21*, 399–404.

[26] Ostojić, B. D.; Stanković, B.; Ðorđević, D. S. The molecular properties of nitrobenzanthrone isomers and their mutagenic activities. *Chemosphere* **2014**, *104*, 228–236.

# [27] Ostojić, B. D.; Stanković, B.; Ðorđević, D. S. Theoretical study of the molecular properties of dimethylanthracenes as properties for the prediction of their biodegradation and mutagenicity. *Chemosphere* 2014, *111*, 114–150.

[28] Stanković, B.; Ostojić, B.; Popović, A.; Gruden, M.; Ðorđević, D. Theoretical study of  
nitrodibenzofurans: A possible relationship between molecular properties and mutagenic  
activity. *Journal of Hazardous Materials* **2016**, *318*, 623–630.

[29] Wilkins, D. M.; Grisafi, A.; Yang, Y.; Lao, K. U.; DiStasio, R. A.; Ceriotti, M. Accurate  
molecular polarizabilities with coupled cluster theory and machine learning. *Proceedings  
of the National Academy of Sciences* **2019**, *116*, 3401–3406.

[30]

## **2. Part II Approach to writing the Paper**

Find out more about the dataset from prof. Stankovic. Find other datasets online that may related to predicting toxicity, carcinogenicity and mutagenicity.

Define search string: AND OR DATE CUT OFF, DOUBLE QUOTES. Search for papers on all three areas: Mutagenicity, Toxicity, Carcinogenicity. With descriptiors like AI, Machine Learning, chemical properties, impact of etc. Find potential gaps in research...are there areas where there appears to be a dearth of research?

If there are areas or combination of areas with gaps in knowledge, formulate general and specific research questions that are relevant to that area.

Synthesize various survey papers to provide more details about the issues not fully covered in other research...this adds value to the knowledge field.

Follow the Survey format from the '1 Dusans\_Reference.pdf' file.

**Reference article**: <https://jcheminf.biomedcentral.com/articles/10.1186/s13321-021-00570-8>

**Chemical toxicity prediction based on semi-supervised learning and graph convolutional neural network**

Toxicity prediction is very important to human and the environment. The toxicity prediction is essential to reduce the cost and labor of a drug’s preclinical and clinical trials. Traditionally, it relies on animal tests, which is time consuming, costly and animals sacrificing (and ethical concerns). Inspired by the success of graph convolutional neural network (GCN) and the needs for improving chemical toxicity prediction confronted with limited data, the authors designed a learning system that hybridizes GCN and semi-supervised learning (SSL) to predict the toxicity of chemical compounds.

In this hybrid learning system, the authors use the Graph convolutional neural network (GCN) to predict chemical toxicity and trained the network by the Mean Teacher (MT) SSL algorithm. In this system, chemical molecules are converted into graphs with one adjacency matrix of dimension N×N and one feature matrix of dimension N×74 (N is the number of atoms in the chemical molecule). After this step, the graph-based molecular data can be learned by the graph convolutional neural network. The architecture of the GCN model consists of two parts, an encoder and a classifier. Bayesian optimization algorithm was used to select the best hyperparameters for constructing the optimal toxicity prediction models.

Certainly, there are merits and drawbacks for this system. One merit is that, unlike others used the entire chemical compound in a one-dimensional feature vector for learning, the proposed system can encode the chemical molecule into a network of features, where the network resembles bond connectivity in the molecule. The system also showed improved prediction on the unlabeled/unannotated data.

The limitation of the system includes:  1) need to increase the diversity of molecular information contained in the feature matrix while limiting the size of the matrix; 2) the interpretability of our graph convolution model has not been explored; 3) the activity cliffs problem has not yet been solved; 4) The impact of different SSL algorithms on the toxicity prediction needs further research.

## **3. Overleaf Document and framework**

**Abstract**

Toxicity prediction of chemical compounds is an important measure in many areas as humans are exposed to an abundance of chemical compounds via the environment, nutrition, cosmetics, and drugs. To protect humans from potentially harmful effects, these chemicals must pass reliable tests for adverse effects and for toxicity. A compound's effects on human health are assessed by a large number of time- and cost-intensive in vivo or in vitro experiments. In particular, numerous methods rely on animal tests, trading off additional safety against ethical concerns.

Lately, there is a growing number of initiatives for computational models to replace biological experiments, but in order for that to happen they must achieve comparable accuracy. In the era of Big Data and Artificial Intelligence, toxicity prediction can benefit from its shift from chemical structural description only to that combined with other available information, which can greatly enhance prediction accuracy. In order to obtain better performance on several toxicity benchmark tasks we developed a computational modeling framework that includes information from various sources 2D and 3D descriptors, and quantum chemistry information (polarizability).

**Introduction**

Chemical toxicity is an important measure in environmental, agricultural, and pharmaceutical science. In the environmental context, toxic chemicals may cause varieties of chronic diseases. In pharmacology, toxicity prediction plays a vital role in the drug discovery pipeline. This makes toxicological screening to be mandatory for the development of new drugs and for the extension of the therapeutic potential of existing molecules.

Several *in vitro/in vivo* techniques have been devised to determine varieties of toxic effect including eye irritancy test, mutagenicity, toxicokinetics, neurotoxicity, embryotoxicity, and genetic toxicity. However, these techniques for examining chemical toxicity are highly cost- and time-intensive. In addition, ethical concerns have also been raised against the usage of animals for toxicology screening.

Therefore, there is an increased demand for cost- and time-efficient toxicological screening methods. *In silico* approach for toxicity prediction has clearly opened a new avenue to address the initial round of screening with acceptable accuracy. These {\em in silico} methods are primarily modeled using high throughput screening assays on various tasks of toxicity. In recent years, Big Data and Artificial Intelligence and in particular machine learning methods have been widely used in toxicity prediction.

However, only chemical structure description has been used to predict toxicity. In order to obtain better performance on several toxicity benchmark tasks we developed a computational modeling framework that includes information from various sources 2D and 3D descriptors, polarizability and quantum chemistry information.

**Background**

Polarizability (α) is linear response of a matter to an external electric field, in which electric  
dipole moment is induced. In general, polarizability is three-dimensional tensor. However, in  
majority of considerations, average of tensor’s diagonal elements is used (α = 1/3 × (αXX +  
αY Y + αZZ )  
Moreover, polarizability is an important property, since it defines the strength of dipole-  
induce dipole and dispersive interactions.  
In their paper 1 stated that contemporary computational research overlooks the importance of twentieth-century findings of electronic interactions between a receptor and a ligand.  
More precisely, most of the current methods for Structure Activity Relationships estimations  
do not take into consideration of polarizability.  
Although1 considered /alpha as a sum of atomic polarizabilities, they showed that it can  
be used for prediction of various types of biological activities, where R2 can be even 0.972.  
However, this approach has two drawbacks, and that is that neither different isomers, nether  
neutral species form ions can be distinguished.  
On the other hand, it is well-known that mutagenic activities of isomers of different  
molecule can be in interval of a few order of magnitudes.2  
Since dipole-induce dipole and dispersive interactions can govern the strength of interac-  
tions between mutagenic molecule and enzyme which activate it, many authors found good  
correlation between mutagenicity of isomers of some molecule and computed polarizabilities.  
Toxicity of chemical species is often modelled using descriptors and molecular fragments.  
The research from the literature often found correlation between size, unsaturation, number  
(a mutual arrangement) of heteroatoms, presence of (cyclic/aromatic) ring systems, presence  
of different molecular groups within the molecule and toxicity. Moreover, these models  
often need to make compromise between simplicity, robustness, and accuracy. This as a  
consequence has usage of 1D and 2D descriptors which looks for the simple and general patterns in the structures.  
Molecular properties calculated using quantum chemistry methods are rarely used as  
descriptors, due to the cost of their calculations. However in the recant years there was  
several attempts to develop ML procedures by which this quantities can be calculated for  
the reasonable amount of time and without (significant) loss of an accuracy. This opens the  
possibility to estimate properties using ML methods and then use these results to predict  
toxicity by another ML algorithm.

## **4 Sim’s Notes:**

Terms to clarify:

Mutagenicity is estimated using the AMES method

IN VIVO experiments

AMES / QSAR challenge

IN SILICO: How does CHEM structure affect Biological interactions

Molecular fingerprints and molecular descriptors

What are ECFP and FCFP fingerprints.

What is Polarizability

 We want to find if there are certain factors (Chemical or otherwise) that can be used to predict toxicity.

Toxicity in context of what? Once it has been ingested? Or something else?

How do we define and measure Toxicity (like AMES for mutagenicity)

Data sets available? Are these Labelled?

**Surveys**

Predicting Mutagenicity from Chemical structure is common place.

13 Molecular properties (MP), ECFP and FCFP fingerprints. 28/6

* Abdul’s code- <https://github.com/Abdulk084/HybridTox2D/blob/master/code.ipynb>

Goal: Show that polarizability contributes to improvement of toxicity prediction (compare with [karim2019efficient (Links to an external site.)](https://pubs.acs.org/doi/abs/10.1021/acsomega.8b03173))

1. Interpretable data driven decision making framework that relies on [wilkins2019accurate (Links to an external site.)](https://www.pnas.org/content/116/9/3401.short)and their application on [https://www.materialscloud.org/work/tools/alphaml (Links to an external site.)](https://www.materialscloud.org/work/tools/alphaml) for modeling polarizability
   * we should think of how to build the software such that we could use other quantum chemistry properties
   * Dataset: [https://archive.materialscloud.org/record/2019.0002/v3 (Links to an external site.)](https://archive.materialscloud.org/record/2019.0002/v3)
   * Another Dataset: [https://www.kaggle.com/zaharch/quantum-machine-9-qm9 (Links to an external site.)](https://www.kaggle.com/zaharch/quantum-machine-9-qm9)
2. Data sources from [chakravarti2019descriptor (Links to an external site.)](https://www.frontiersin.org/articles/10.3389/frai.2019.00017/full): [Mutagenicity\_N6512.csv](https://psu.instructure.com/courses/2174348/files/128993454?wrap=1) [Download Mutagenicity\_N6512.csv](https://psu.instructure.com/courses/2174348/files/128993454/download?download_frd=1)

Please, ask for access to this Colab where Dr. Stankovic played with the files:

[https://drive.google.com/file/u/0/d/1Wct3Zu3M69VFkdj8JHGUrsv7Q8orNauF/edit (Links to an external site.)](https://drive.google.com/file/u/0/d/1Wct3Zu3M69VFkdj8JHGUrsv7Q8orNauF/edit)

If that doesn't work (it didn't work for me), you can have the file:

[Descriptor.ipynb](https://psu.instructure.com/courses/2174348/files/128995910?wrap=1) [Download Descriptor.ipynb](https://psu.instructure.com/courses/2174348/files/128995910/download?download_frd=1)

Think about this:

[https://www.technologyreview.com/2020/11/30/1012712/deepmind-protein-folding-ai-solved-biology-science-drugs-disease (Links to an external site.)](https://www.technologyreview.com/2020/11/30/1012712/deepmind-protein-folding-ai-solved-biology-science-drugs-disease)