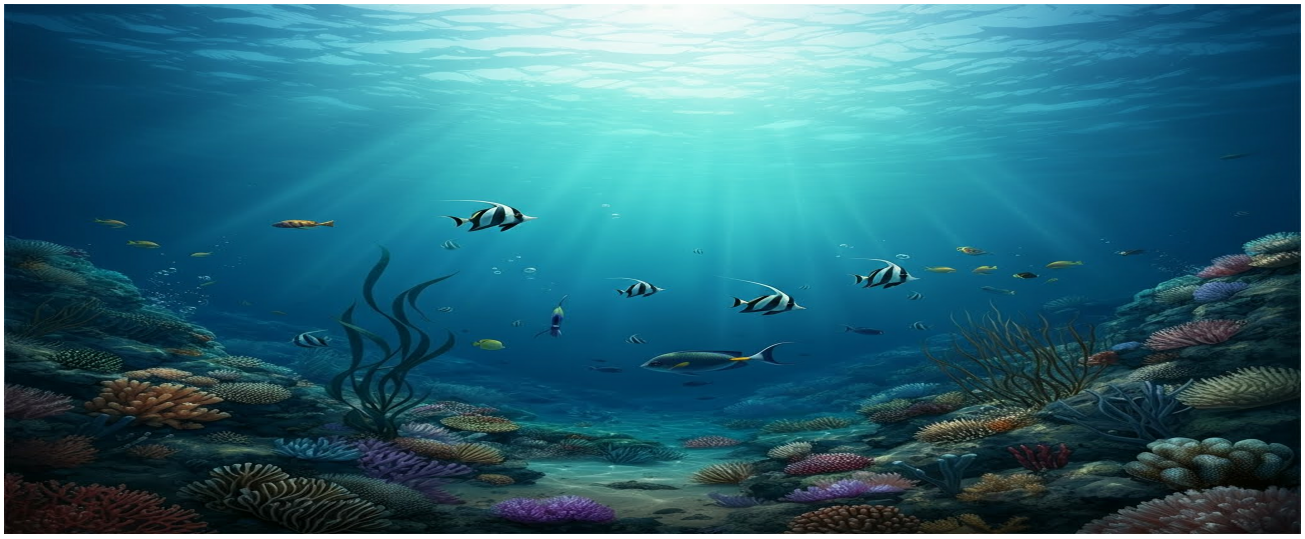
Internship Report

New Alternate Methods for Acute Fish Toxicity Testing: RTGill versus FET and Building Machine Learning Models

July 14, 2025



Author : Soumodeep HOODATY

Promotion : 2025

L'enseignant Referent IP Paris : Clément REY

Supervisor : Yannick BAYONA

Effective Internship Dates: 03/02/2025 to 31/07/2025
Name of Organization : L'Oréal Groupe
Address : 9 Rue Pierre Dreyfus, 92110 Clichy

Contents

Acknowledgement

First and foremost, I extend my sincerest appreciation to my supervisor, Dr. Yannick Bayona, for granting me this incredible opportunity to work under his expert guidance. His unwavering support and prompt responses to my inquiries were instrumental to my progress and learning throughout this internship. I am immensely grateful for his mentorship and the trust he placed in me.

I would also like to express my heartfelt thanks to every member of the Environmental Safety team. Their generosity in sharing their time and expertise in answering my questions, created a supportive and encouraging learning environment. Their willingness to guide me whenever I needed assistance was essential to my understanding the tools available in this domain, and the challenges they overcome to continue doing the excellent work they do.

Furthermore, I wish to thank Dr. Leopold Carron. His assistance with ideation and his valuable and insightful responses to my data and machine learning-related queries and guiding me throughout my journey greatly assisted in the successful completion of my project.

I am truly thankful for the opportunity to have learned from and collaborated with such a dedicated and talented group of individuals. This internship experience has been profoundly impactful, shaping my understanding of environmental safety and providing me with invaluable skills and knowledge that I will carry forward in my career.

1 Abstract

The replacement of animal testing for the environmental risk assessment of cosmetic ingredients is a key scientific and ethical priority. Acute fish ecotoxicity is a regulatory endpoint that has traditionally relied on animal studies, creating a critical need for validated alternative methods. This internship project addresses this need by navigating the complex landscape of non-animal alternatives through two primary objectives. The study follows a dual strategy: first, it evaluates and compares the predictive performance of two prominent non-animal assays, the in vitro RTGill-W1 and the in vivo Fish Embryo Toxicity (FET) tests, within a Defined Approach (DA). Second, it focuses on the development of novel machine learning models from the ground up, using molecular descriptors and fingerprints. Our results show that while a general Random Forest model can achieve a baseline accuracy a far more effective strategy is the creation of specialized models based on chemical clustering. By segmenting chemicals into distinct groups using the HDBSCAN algorithm, the resulting cluster-specific models were able to reduce prediction errors for chemicals within their applicability domain. This research demonstrates that a tailored, cluster-based modelling framework represents a significant step forward in developing accurate and reliable in-silico tools to replace animal testing in ecotoxicology.

2 Computing Specifications

1. Google Cloud Platform

- RAM: 24 Gb
- Processor: No explicit specification, 6 CPUs were used for computation

2. Local Machine

- RAM: 16 Gb
- Processor: 11th Gen Intel (R) Core (TM) i5-1145G7@2.60GHz

3 Definitions

- **Globally Harmonized System (GHS) Classification**: The GHS is an internationally agreed-upon standard managed by the United Nations that was set up to replace the assortment of hazardous material classification and labelling schemes previously used around the world. Core elements of the GHS include standardized hazard testing criteria, universal warning pictograms, and harmonized safety data sheets which provide users of dangerous goods with a host of information. The system acts as a complement to the UN Numbered system of regulated hazardous material transport. Implementation is managed through the UN Secretariat. Although adoption has taken time, as of 2017, the system has been enacted to significant extents in most major countries of the world. This includes the European Union, which has implemented the United Nations' GHS into EU law as the CLP Regulation, and United States Occupational Safety and Health Administration standards. According to United Nations, 2019, ST/SG/AC.10/30/Rev.8.
- **Ecotoxicity**: Ecotoxicity refers to any harmful effect of a chemical substance on living organisms other than human. These effects can affect the survival, growth, development or reproduction of these organisms. It is assessed using standardized tests carried out on species representative of the different links in the food chain of natural ecosystems, whether aquatic or terrestrial. Trigger values based on GHS 09 labelling EU rules.
- **Acute Effect**: A rapidly developing effect caused by a single or brief exposure.

- **LC50**: Lethal effect concentration on 50% of the population.
- **End of Life**: In Life Cycle Assessment, it is the step of the whole life cycle of a product concerning the fate of product after its dedicated use. In L'Oréal, it is an assessment performed to assess the potential amount of substance that can be released into the environment and its impact on the different compartments (air, surface water, ground water, soil) after use. The criteria used to assess the environmental impact and end of life of substances are the persistence, mobility, bio-accumulation and eco-toxicity.
- **Quantitative Structure Activity Relationship (QSAR)**: A qualitative association between a chemical substructure and the potential of a chemical containing the substructure to exhibit a certain biological effect.
- **Simplified Molecular Input Line Entry System (SMILES)**: A simplified chemical notation that allows a user to represent a two-dimensional chemical structure in linear textual form for easy entry into a computer application.
- **Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH)**: It is a regulation of the European Union, adopted to improve the protection of human health and the environment from the risks that can be posed by chemicals, while enhancing the competitiveness of the EU chemicals industry. It also promotes alternative methods for the hazard assessment of substances in order to reduce the number of tests on animals. According to (EC) No 1907/2006.
- **OECD**: The Organisation for Economic Co-operation and Development. It is an intergovernmental economic organization that works to stimulate economic progress and world trade. It facilitates international cooperation in the development and harmonization of approaches for assessing the safety of chemicals, by reducing the need for extensive animal testing and streamlining regulatory processes across member countries.
- **OECD QSAR Toolbox**: The OECD QSAR Toolbox is a regulatory software platform that facilitates the development of quantitative structure-activity relationships for chemical safety assessment and supports data gap filling through chemical grouping and read-across approaches.
- **CAS Number**: A CAS Registry Number (also referred to as CAS RN or informally CAS Number) is a unique identification number, assigned by the Chemical Abstracts Service (CAS) in the US to every chemical substance described in the open scientific literature, in order to index the substance in the CAS Registry.
- **Sensitivity**: $\frac{\text{True Positive}}{\text{True Positive + False Negative}}$
- **Specificity**: $\frac{\text{True Negative}}{\text{True Negative + False Positive}}$
- **Balanced Accuracy**: $\frac{\text{Sensitivity + Specificity}}{2}$
- **RTGill-W1 assay**: This assay is a standard assay registered as the OECD 249 guideline[1]. The RTgill-W1 cell line assay describes a 24-well plate format fish cell line acute toxicity test using the permanent cell line from rainbow trout (Oncorhynchus mykiss) gill, RTgill-W1. After 24 hours of exposure to the test chemical, cell viability is assessed based on three fluorescent cell viability indicator dyes, measured on the same set of cells. Resazurin enters the cells in its non-fluorescent form and is converted to the fluorescent product, resorufin, by mitochondrial, microsomal or cytoplasmic oxidoreductases. A reduction in the fluorescence of resorufin indicates a decline in cellular metabolic activity, including disruption of mitochondrial membranes. The data are expressed as the percent cell viability of unexposed control values versus the test chemical concentration.

- **FET (Fish Embryo Toxicity) test**: This assay is a standard assay registered as the OECD 236 guideline. The test method is intended to determine the acute or lethal toxicity of chemicals on embryonic stages of fish (Danio rerio). Newly fertilised zebrafish eggs are exposed to the test chemical for a period of 96 hours, every 24 hours. Twenty embryos (one embryo per well) are exposed to the chemical tested at each concentration level. The test includes five increasing concentrations of the chemical tested and a control. Every 24 hours, four apical observations are recorded as indicators of lethality:
  - coagulation of fertilised eggs
  - lack of somite formation
  - lack of detachment of the tail-bud from the yolk sac
  - lack of heartbeat

  At the end of the exposure period, acute toxicity is determined based on a positive outcome in any of the four apical observations recorded and the LC50 value is calculated.
- **In-silico**: In biology and other experimental sciences, an in-silico experiment is one performed on a computer or via computer simulation software.
- **In-vitro**: In-vitro studies are performed with microorganisms, cells, or biological molecules outside their normal biological context. Colloquially called ”test-tube experiments”, these studies in biology and its subdisciplines are traditionally done in lab-ware such as test tubes, flasks, Petri dishes, and microtiter plates. Studies conducted using components of an organism that have been isolated from their usual biological surroundings permit a more detailed or more convenient analysis than can be done with whole organisms; however, results obtained from in-vitro experiments may not fully or accurately predict the effects on a whole organism.
- **In-vivo**: In vivo studies are those in which the effects of various biological entities are tested on whole living organisms or cells, usually animals, including humans, and plants, as opposed to a tissue extract or dead organism.

4 Introduction

4.1 L'Oréal & Environmental Safety

L'Oréal Groupe is the world's largest cosmetics company, a French multinational personal care corporation founded in 1909 by chemist Eugène Schueller. What started with a hair dye formula has evolved into a global leader in the beauty industry, marketing 37 global brands across four divisions: Professional Products, Consumer Products, Luxe, and Dermatological Beauty. Sustainability is a core pillar for L'Oréal, embodied in its "L'Oréal for the Future" commitments, focusing on reducing environmental impact. L'Oréal is also recognized for its early stance against animal testing. According to its commitments, L'Oréal Groupe conducts a strict environmental impact assessment on its products and ingredients used in formulas. This evaluation is carried out in several stages based on data from the regulatory dossier, scientific literature, from modelling or from specific testing.

For each ingredient, L'Oréal Groupe analyses the environmental data publicly availables and those provided by its suppliers. The company's environmental security evaluators look for additional information in the scientific literature and international regulatory databases. When the chemical structure of the ingredients is well define, safety evaluators can also carry out computer modelling of their physicochemical properties, ecotoxicity, mode of action and biodegradability. As the company is responsible of safety its formulas, further testing may be performed to fulfill the knowledge gaps identified.

The environmental adventure start in 1995, when L'Oréal Groupe created its Environmental Advanced Research Laboratory whose mission is investigate an develop new alternative methods to evaluate the impact of ingredients and formulas on organisms other than human. The company is working on improving the environmental profile of formulas to improve their environmental impact. As the regulation constraint growth, in 2017, L'Oréal Groupe created its Environmental Safety teams involved in the overall assessment and defense of ingredients. This teams is working to define science based strategy of defense, ensure the safety watch of portfolio ingredient and ensure the screening of new-coming ingredients.

4.2 Global context

In the context of the cosmetic regulation, which only regulates human safety, the cosmetics industry has made significant progress in eliminating animal testing. Other regulations such the European Chemical Regulation for Chemicals (REACH), or Chinese Chemical Regulation investigate both human and environmental safety through international standard methods involving vertebrate models in standard tests. According to the 3R (**R**eplace animal experiments wherever possible, **R**educe the number of animals used, **R**efine experiments to minimise the effect on animals) objectives and global reduction of animal testing in line with the EU Directive on the protection of animals used for scientific purposes (EU, 2010)[2], the environmental safety and professional associations aimed to develop new alternative methods to ensure under regulations such as the European Union's REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals), environmental safety data was mandatory to evaluate how chemical ingredients impact aquatic ecosystems and biodiversity. The assessment strategy aimed to cover three trophic levels of aquatic organisms, including algae, invertebrates, and fish, investigating both acute and chronic effects as well as bioaccumulative potential.

In Vivo Approaches: The Traditional Standard The environmental safety assessment relies on *in vivo* testing, involving exposure of living fish species such as rainbow trout, fathead minnows, or zebrafish to various concentrations of chemical substances. These tests, performed on young or adult living organisms, represent the current regulatory gold standard but create a complex regulatory paradox: while a cosmetic ingredient may no longer require animal testing for human safety purposes under cosmetic regulations, it might still require animal testing to meet human and environmental safety

requirements under other regulatory frameworks.
In Vitro Methods: Cell-Based Alternatives The limitations and ethical concerns surrounding fish testing have catalyzed intensive research into *in vitro* alternatives. These approaches utilize embryo models or fish cell lines derived from species of interest, offering a bridge between traditional animal testing and computational methods while maintaining biological relevance to aquatic organisms.
In Silico Solutions: Computational Modelling The most promising long-term solutions lie in *in-silico* approaches, encompassing modeling techniques and machine learning algorithms that can predict environmental toxicity without biological testing. These computational methods represent the ultimate goal of completely eliminating animal testing from chemical ingredient safety assessment.

The Global Harmonized System of Classification and Labelling of Chemicals (GHS) was developed by the United Nations to provide a unified framework for chemical hazard communication worldwide. The European Union implemented GHS through the CLP Regulation (Classification, Labelling and Packaging), which ensures consistent hazard identification and communication across member states.

Figure 1: *GHS09 Environmental Hazard Pictogram - indicating substances hazardous to the aquatic environment*

The GHS09 pictogram (Figure 1) features a dead fish above a dead tree, symbolizing acute and long-term environmental damage to aquatic ecosystems. This pictogram is assigned to substances that meet specific environmental hazard criteria.

Categories	**Acute Aquatic Toxicity**	**Chronic Aquatic Toxicity**
Category 1:	LC50/EC50 $\leq$ 1 mg/L (most toxic)	NOEC $\leq$ 0.1 mg/L (most concerning)
Category 2:	1 $<$ LC50/EC50 $\leq$ 10 mg/L	0.1 $<$ NOEC $\leq$ 1 mg/L
Category 3:	10 $<$ LC50/EC50 $\leq$ 100 mg/L	1 $<$ NOEC $\leq$ 10 mg/L
Category 4:	–	10 $<$ NOEC $\leq$ 100 mg/L

Table 1: *GHS Classification categories based on concentration-based ecotoxicity values*

Here EC50 (EC–Effect Concentration), similar to LC50, is the concentration of a chemical at which 50% of its maximum response is observed, while NOEC (No Observed Effect Concentration) is the the highest concentration tested at which no effect on the living organism was observed. **Acute aquatic toxicity** refers to adverse effects occurring within short-term exposure periods (typically 96 hours for fish, 48 hours for invertebrates, 72 hours for algae), measured as LC50/EC50 values, while **chronic aquatic toxicity** encompasses long-term effects resulting from prolonged or repeated exposure over extended periods (21-28 days or longer), assessed through NOEC values that indicate the highest concentration causing no statistically significant adverse effects. These trigger values determine the environmental hazard classification and labelling requirements, with Category 1 and 2 substances requiring the GHS09 pictogram and specific hazard statements such as ”Very toxic to aquatic life” or both ”Very toxic” or ”Toxic to aquatic life with long lasting effects.” All our results are based on data for acute fish toxicity which follow the same trigger values (LC50).

The industry is at a critical juncture where the complete elimination of animal testing depends largely

on developing and validating robust alternatives for environmental safety assessment. Recent initiatives, such as the EU's 2024 roadmap[3] to phase out all animal testing for chemical safety assessments and similar commitments from other regulatory bodies, signal a growing recognition that the future of chemical and cosmetic safety evaluation lies in innovative, animal-free methodologies.

This evolution represents not just an ethical imperative but also a scientific opportunity to develop environmentally predictive safety assessment tools that can better protect ecosystems.

5 Materials and Methods

5.1 Chemical Data Processing Libraries

PubChemPy is a Python wrapper that provides programmatic access to the PubChem database, one of the world's largest collections of freely accessible chemical information.

RDKit (RD stands for Rational Discovery) is an open-source cheminformatics and machine learning software toolkit written in C++ with Python bindings. RDKit excels at generating molecular fingerprints, computing physicochemical properties, performing substructure searches, and handling chemical transformations, making it an essential tool for drug discovery and computational chemistry workflows.

Chem (specifically `rdkit.Chem`) is the core molecular handling module within RDKit that provides fundamental classes and functions for molecular representation and manipulation. This module serves as the foundation for most RDKit operations and molecular data processing tasks.

These packages collectively provide a robust framework for computational chemistry research, enabling efficient handling of large-scale chemical datasets and molecular analysis workflows.

5.2 Defined Approach and Silico Models Selection

Based on *MacMillan et al.* [4], the first step was a comprehensive understanding of the underlying tools and reproduction of methods used in the defined approach proposed. It is highly important to note that the authors of most of the models used do not provide any meta-data regarding the molecular descriptors, chemical fingerprints or the data that were used to create them. The data set which is used for testing and benchmarking of all the models is available in the supplementary information of the reference paper. The KATE model (KAshinhou Tool for Ecotoxicity) was implemented according to the paper, *Zhou et al.* [5] as the paper gave strong evidence for good classification power.

The models would either predict the LC50 values of the chemicals or predict the GHS classification of the chemical:

Prediction and Applicability Domain analysis for model:
Fish Acute (LC50) Toxicity classification (SarPy-IRFMN) (version 1.0.3)
(calculation core version : 1.3.19)

No.	Id	SMILES	Assessment	Predicted toxicity class	Experimental	Structural Alerts	ADI	Similarity index	Accuracy index	Concordance index	ACF index	Remarks
1	Molecule 1	O=[N+]([O-])c1ccc(O)cc1	Toxic-3 (between 10 and 100 mg/l) (EXPERIMENTAL value)	Toxic-3 (between 10 and 100 mg/l)	Toxic-3 (between 10 and 100 mg/l)	Toxicity class 3 alert no. 13; Toxicity class 3 alert no. 24	1	1	1	1	1	-
2	Molecule 2	c1ccc(cc1)NC	NON-Toxic (more than 100 mg/l) (EXPERIMENTAL value)	Toxic-3 (between 10 and 100 mg/l)	NON-Toxic (more than 100 mg/l)	Toxicity class 3 alert no. 24	1	1	1	1	1	-
3	Molecule 3	NNc1ccccc1	Toxic-3 (between 10 and 100 mg/l) (LOW reliability)	Toxic-3 (between 10 and 100 mg/l)	-	Toxicity class 3 alert no. 24	0.474	0.865	1	1	0.51	-

Figure 2: *VEGA SARPY-IRFMN Model (Classification model based on GHS system)*

Prediction and Applicability Domain analysis for model:
Guppy LC50 model (KNN-IRFMN) (version 1.1.2)
(calculation core version : 1.3.19)

No.	Id	SMILES	Assessment	Predicted toxicity [log 1/LC50(mmol/L)]		Predicted toxicity [mg/l]		Molecules used for prediction		Molecular						
Weight	Experimental value [mg/l]			Experimental	ADI	Similarity index		Accuracy index		Concordance index		Max error index		ACF index		
	Remarks															
1	Molecule 1	O=Cc1ccc(O)c(OCC)c1		84.99 mg/L (LOW reliability)		0.29	84.99	2	166.19	-	-	0.532	0.887	0.206	0.2	
	0.4	0.6	-													
2	Molecule 2	O=Cc1ccc(O)c(OC)c1		80.28 mg/L (LOW reliability)		0.28	80.28	2	152.16	-	-	0.527	0.879	0.054		
	0.215	0.096	0.6	-												
3	Molecule 3	O=C(OCC)CC(C(=O)OCC)SP(OC)(OC)=S		N/A	-	-	0	330.4	-	-	-	-	-	-	-	-
	-	[Model]Unable to perform Applicability Domain check.														

Figure 3: *VEGA GUPPY Model (Regression model—later converted to GHS classes)*

Regardless of the model, the predictions were converted into the required classification based on the LC50 value or class predicted. The selection of the relevant models were done on the basis of the method laid down in the reference pa and definitionper. We first note the sensitivity and specificity of each of the models tested using the actual LC50 values from the reference data, which then helps us calculate the balanced accuracy of each of the models.

Using the balanced accuracy performance, we select the **top 3** models from our earlier selection — based on various state-of-the-art models that are currently used all across the globe for toxicity prediction. This is then used in the defined approach. It is important to note that we created a Python function to remove all the predictions with low reliabilities. The reliability of the prediction is mentioned along with the prediction itself (see Figure 2 and 3). This approach is applied to replicate the steps undertaken in the reference paper.

5.3 Mechanism of Action

Mechanism of action (MechoA)[6] refers to the molecular interaction that a molecule will undergo, leading to a biological outcome which can be the key starting point of the Adverse Outcome Pathway (AOP) for this substance, i.e. the Molecular Initiating Event (*Allen et al.*[7]).

Mechanism of action provides crucial data that might turn out to be useful for building models around toxicity prediction. We can define the mechanism of action into several different classes:

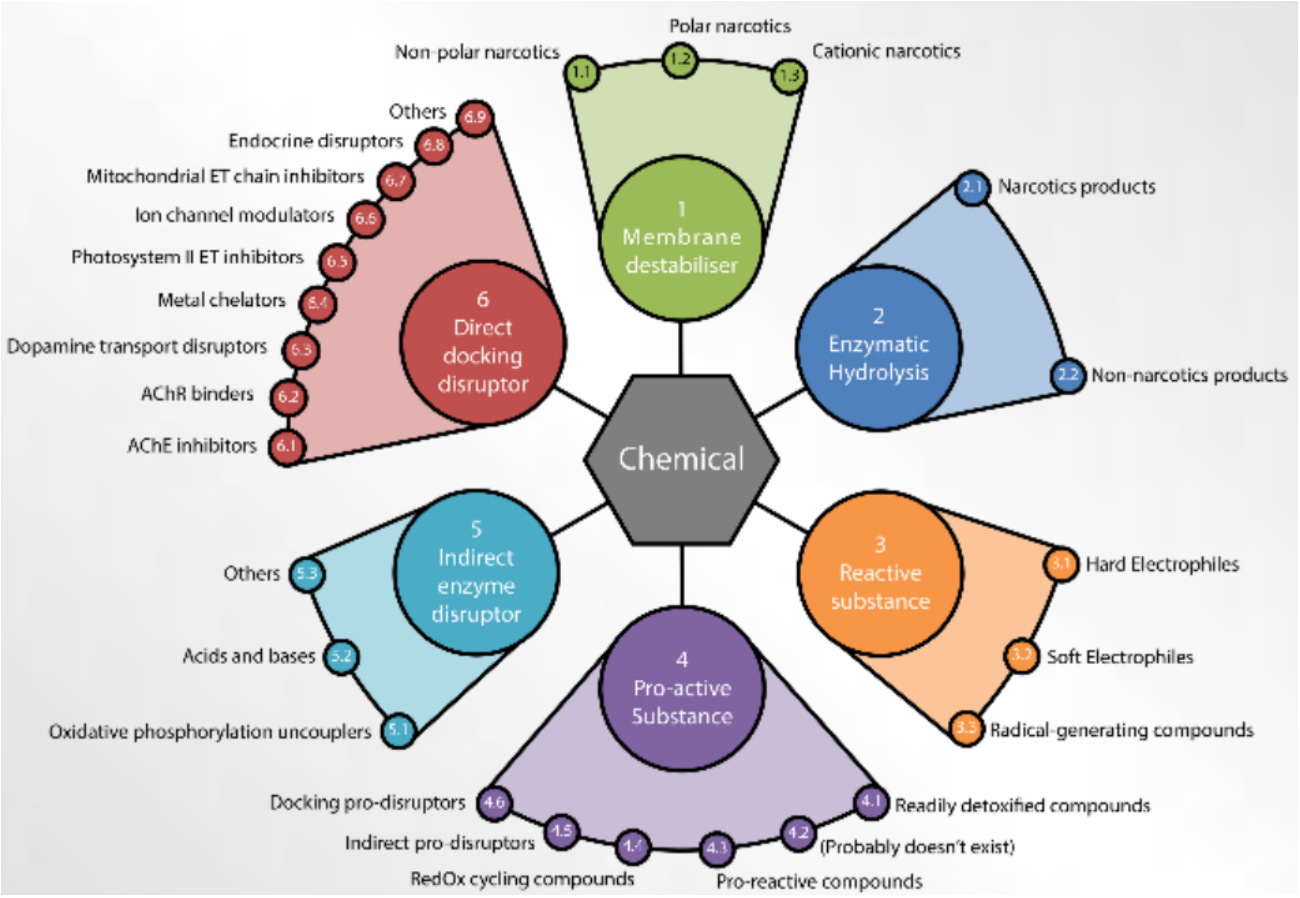


Figure 4: *Chart of the different MechoA profiles[8]*

The classification of the MechoA profiles of the training set into the 6 relevant broader classes was done using the iSafeRat Profiler[9] available as a plug-in on the OECD QSAR Toolbox[10].

5.4 Defined Approach

A defined approach (DA) [11] to testing and assessment consists of a fixed data interpretation procedure (DIP) applied to data generated with a defined set of information sources to derive a result that can either be used on its own, or together with other information sources within an Integrated Approach to Testing and Assessment (IATA), to satisfy a specific regulatory need. Thus, a defined approach to testing and assessment can be used to support the hazard identification, hazard characterisation and/or safety assessment of chemicals.

The defined approach in our context, is a step-by-step method which is used to improve the efficiency and overall accuracy of the GHS class prediction. In its essence, it is a scoring method that is used to rate the prediction of each of the individual models:

GHS Class	Score
Acute 1 (0-1 mg/L)	4
Acute 2 (1-10 mg/L)	2
Acute 3 (10-100 mg/L)	1
Non-Toxic (>100 mg/L)	0

Table 2: *GHS Classification and Scoring System*

The first step was to check the chemical with the RTGill-W1 assay (from hereon will be written as simply RTGill) based on the method described in MacMillan et al. [4], run their predictions individually for each model and proceed with the defined approach method. In terms of predictivity, RTGill by itself provides an accuracy of 50% in predicting the experimental LC50 values in the reference data. In total, there are 66 chemicals for which RTGill data is available which significantly reduces the data on which we can analyse the results of our models. We then score the predictions of the top 3 models which in our case are:

- **VEGA Fathead KNN [12]:** The VEGA Fathead Minnow KNN model uses a k-Nearest Neighbor algorithm (k=4) to predict acute lethal toxicity (LC50) in fathead minnow fish. The model identifies the four most structurally similar compounds from its training dataset using a similarity index, then calculates the prediction as a weighted average of these neighbors' experimental toxicity values. Unlike traditional QSAR models, it relies on structural similarity rather than molecular descriptors for predictions. The model includes an Applicability Domain assessment that provides reliability scores (good: $>$0.75, moderate: 0.7-0.75, low: $<$0.7) to indicate prediction confidence.

- **VEGA Fish KNN [13]:** The model performs a read-across on a dataset of 972 chemicals. This dataset has been made by Istituto di Ricerche Farmacologiche Mario Negri. The read-across model has been built with the istKNN application (developed by Kode srl, http://chm.kodesolutions.net) and it is based on the similarity index developed inside the VEGA platform; the index takes into account several structural aspects of the compounds, such as their fingerprint, the number of atoms, of cycles, of heteroatoms, of halogen atoms, and of particular fragments (such as nitro groups). The index value ranges from 1 (maximum similarity) to 0. On the basis of the structural similarity index, the four compounds from the dataset resulting most similar to the chemical to be predicted are taken into account; compounds with a similarity value lower than 0.7 are discarded, and if only one compound remains available for prediction, it is kept only if it has a similarity value higher than 0.75. If no compounds fall under these conditions, no prediction is provided. The estimated toxicity value is calculated as the weighted average value of the experimental values of the chosen compounds, using their similarity values as weight. Their similarity values are raised to the power of 3 in order to enhance the weight of the most similar compounds in the calculated prediction.

- **ECOSAR Freshwater Model [14]:** The model was developed by classification/sub-classification of chemicals based on similarity of structure and similarity in measured effect levels from aquatic toxicity data. When available, modes-of-action have been integrated into the classification scheme to substantiate trends seen in available data. Reliable predictions are made by the model when the log Kow values are $\leq$5.0. U.S. EPA has focused resources on models for aquatic toxicity to freshwater organisms because most releases of industrial chemicals go to freshwater bodies. There is no other insight or document that could be found that could give us a better picture about the algorithm of the model.

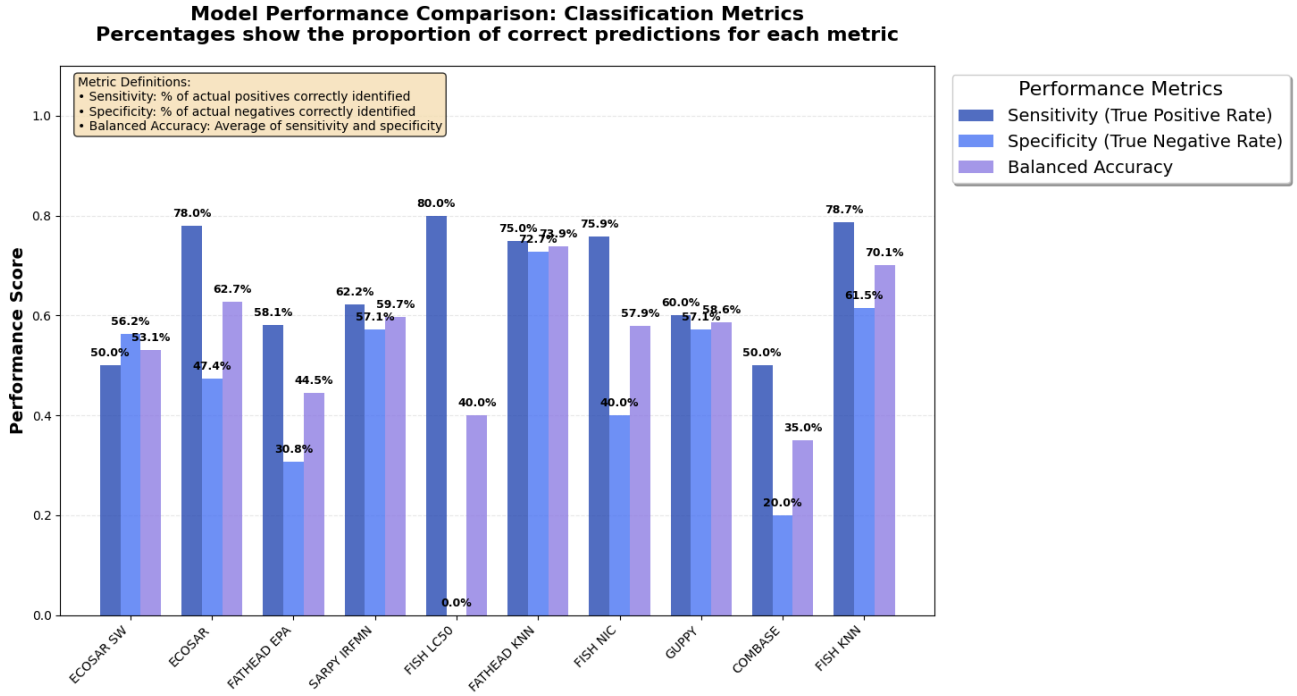


Figure 5: *Sensitivities, specification and balanced accuracies of predictions of all models tested with respect to the actual LC50 values*

There are two methods which can be employed for the implementation of the defined approach. We can either consider to take into account the chemicals which have in-domain predictions from all the 3 models, or those that have predictions from 2 out of the 3 models. We also decided to implement the KATE acute toxicity for fish model as mentioned earlier (since its performance matches that of ECOSAR in terms of balanced accuracy and provides higher sensitivity).

After the scoring is complete, the average score is calculated for each prediction using this approach, followed by rounding them off. Using the average rounded off score, we can reverse map the scores to their respective GHS class, which is thus the prediction given by the defined approach.

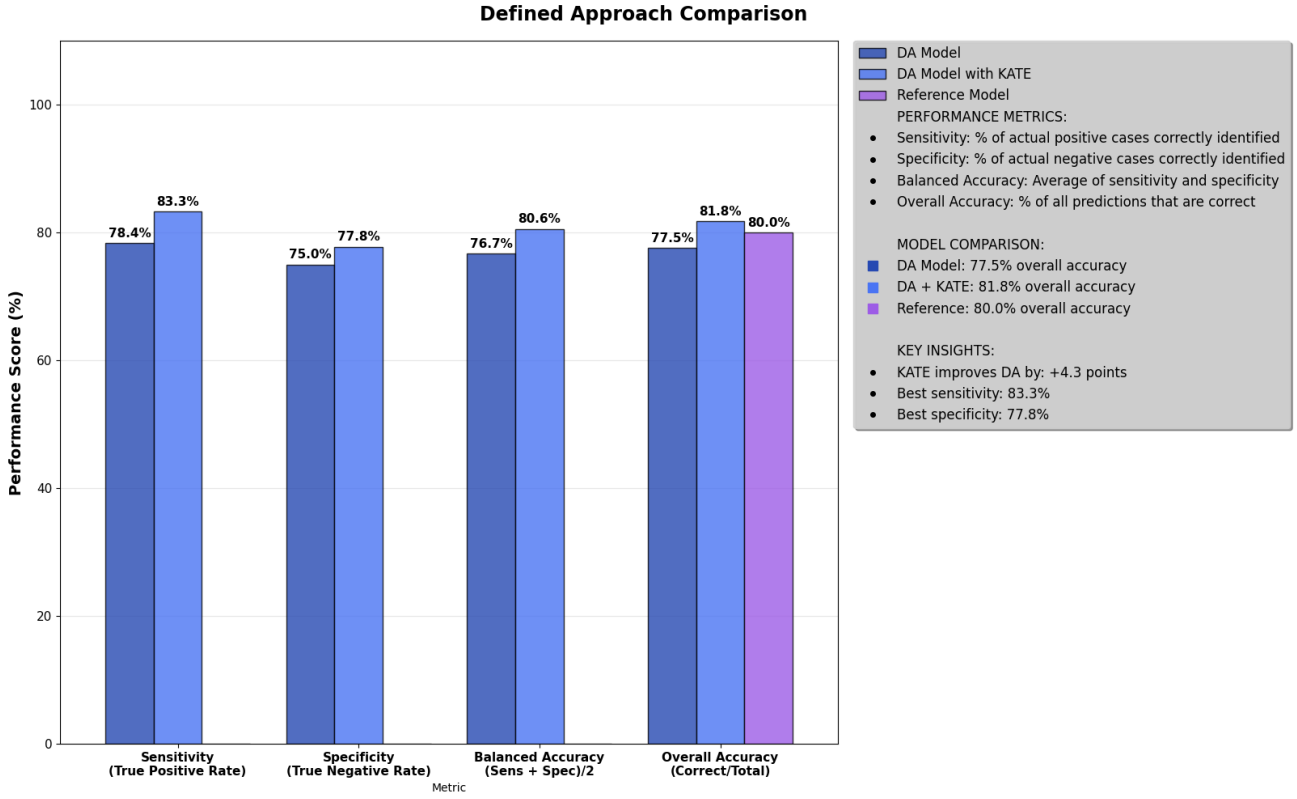


Figure 6: *Comparison between published reference accuracy, calculated DA accuracy, KATE included DA accuracy*

The calculated accuracy of the DA without KATE, as done in the reference paper is less than the value of 80% mentioned in the paper, which may be due to the usage of different versions of the toolbox (version 4.5 used in the paper as compared to version 4.7 for our calculations) leading to different in-domain predictions. Regardless, overall accuracy improves and is better than the best individual model in every aspect. An interesting fact to note is that the specificity which was quite inconsistent with the individual models seem to have improved as well with this approach. Also, the effects of implementing the KATE model in the defined approach can also be seen, as noteworthy improvements in results can be clearly observed. However, it is highly essential to note that the number of in-domain chemicals are relatively lower when KATE is added to the defined approach which might contribute to its better performance to some extent as there are lesser chemicals to evaluate (36 instead of 49). Regardless, this does not take away from the fact that the defined approach benefits from adding or modifying the models that were initially used.

This was followed by the implementation of FET data instead of RTGill to explore the possibilities of performance changes. There are a total of 150 chemicals for which FET data is available, of which 147 of them have LC50 values in our experimental data. The majority of comparisons and suggestions are restricted based on only the common chemicals for which RTGill and FET data are available (21 chemicals in total). By itself, FET provides an accuracy of 58.5% on chemicals from the reference data. When used in the DA as a replacement for the RTGill data for both the DA with and without the KATE model, it consists of only 18 and 12 chemicals for which it has the data, respectively, providing an accuracy of 66.67% for both. Next, we looked at the per-class accuracy of both tests to understand the differences in predictivity between them.

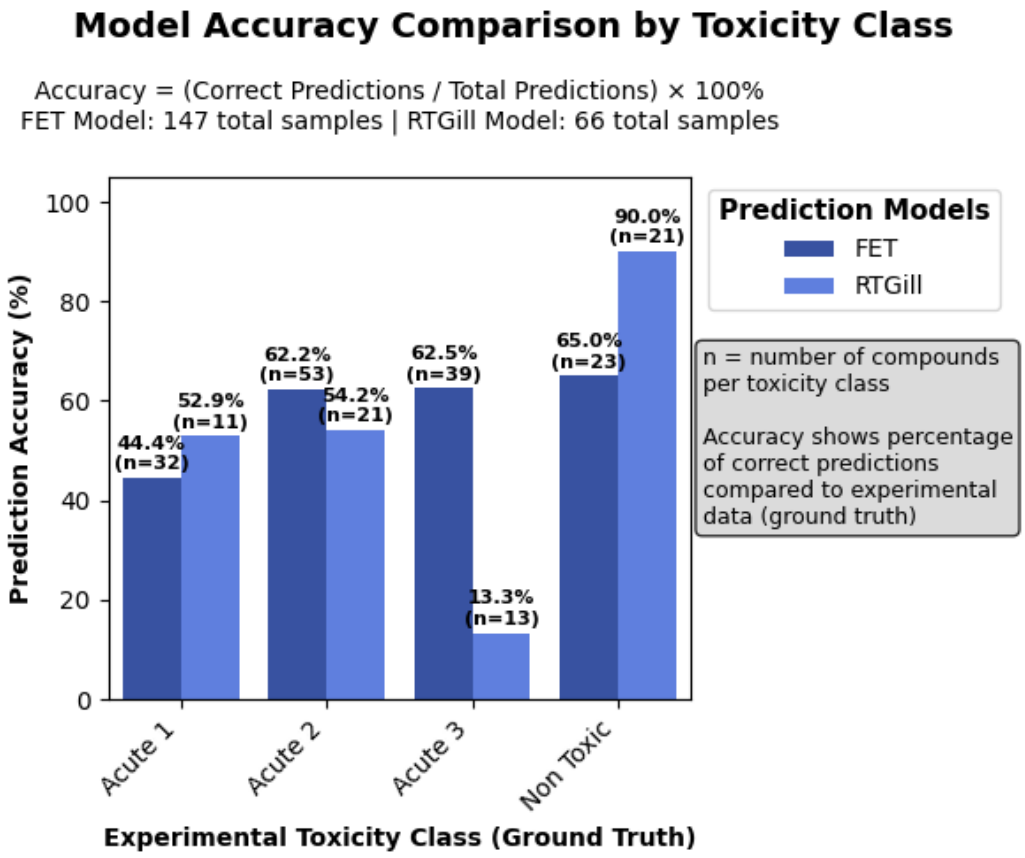


Figure 7: *Comparison between FET and RTGill data prediction accuracy*

Comparative Analysis

Looking at the first 3 classes in Figure 6, Acute 1, 2 and 3, we can observe that the availability of data is generally more than double for FET as compared to RTGill, while for non-toxic compounds, it is almost the same number of compounds. It is interesting to note, that even with a higher number of compounds for FET data, it is providing us with a higher accuracy for the acute classes 2 and 3, with FET providing exceptionally better accuracy with triple the number of compounds for the acute 3 class of chemicals. For acute 1, even though the accuracy is lower, the number of chemicals used to calculate it is almost three times that of RTGill. However, RTGill experiments seem to provide a higher percentage of accuracy in identifying non toxic compounds as compared to FET with a similar number of compounds.

We decided to test whether using both FET data and RTGill data would aid in making a more informed decision in terms of accuracy, however it rather reduced the accuracy of the DA to 61.11% with 18 chemicals and of the DA with KATE to 58.33% with 12 chemicals.
After further investigation, we found that for the 21 compounds in common between RTGill and FET data, there were 5 cases in which both predictions were incorrect, 8 cases where both provide the correct classification, while the rest 8 have a mismatch in classification. This results in reinforcing the error classifications and reducing the impact of correct predictions in the DA.

Further statistical tests were carried out based on the data available. It should be kept in mind that we are performing our analysis on only 21 overlapping compounds between FET and RTGill data, and that is being done to take the initiative of making more calculated and data-driven decisions in the future. The **Cohen's kappa** was calculated, and we found that the agreement between the experimental data and FET data had a score of 0.436 (moderate), which is slightly higher than RTGill data and the experimental data, which is 0.332 (fair). FET data and RTGill data had a score of 0.352 (fair) between

them. This suggests that FET values align better with the experimental LC50 values as compared to the RTGill test, however our data is too small for us to be able to verify this hypothesis.

Because of our very small sample size, we decided to use the **Wilson confidence interval** for each class for both of the tests to check the range in which the accuracy for each test per-class lies, and also for the overall prediction with 95% confidence:

Method	Overall	Acute 1	Acute 2	Acute 3	Non Toxic
RTGill	[0.324, 0.717]	[0.158, 0.750]	[0.250, 0.842]	—	[0.376, 0.964]
FET	[0.409, 0.792]	[0.250, 0.842]	[0.250, 0.842]	—	[0.376, 0.964]

Note: Values show accuracy (point estimate) and [Wilson 95% CI]. Acute 3 CIs not calculated due to presence of only 2 samples in the overlapping data.

Table 3: *Method Accuracy with Wilson 95% Confidence Intervals by Toxicity Class*

From table 2, it is clear that all the intervals that were calculated are **quite wide** indicating low precision. For example, for the RTGill test, the accuracy for predicting 'Acute 1' chemicals with 95% confidence, lies between 15.8% and 75%. This suggests that there is not enough data to narrow down the range and hence fails to inform us about which test may be better.

From these tests we can easily see the need of more data on the tests being a necessity to have a more statistically significant analysis and conclusion regarding their comparison. Hence, for a better idea of the sample size that may be required to conduct a detailed and rigorous analysis with a certain amount of confidence, we assume a normal distribution and estimate it, and the calculations related to it can be found in section 8.1.

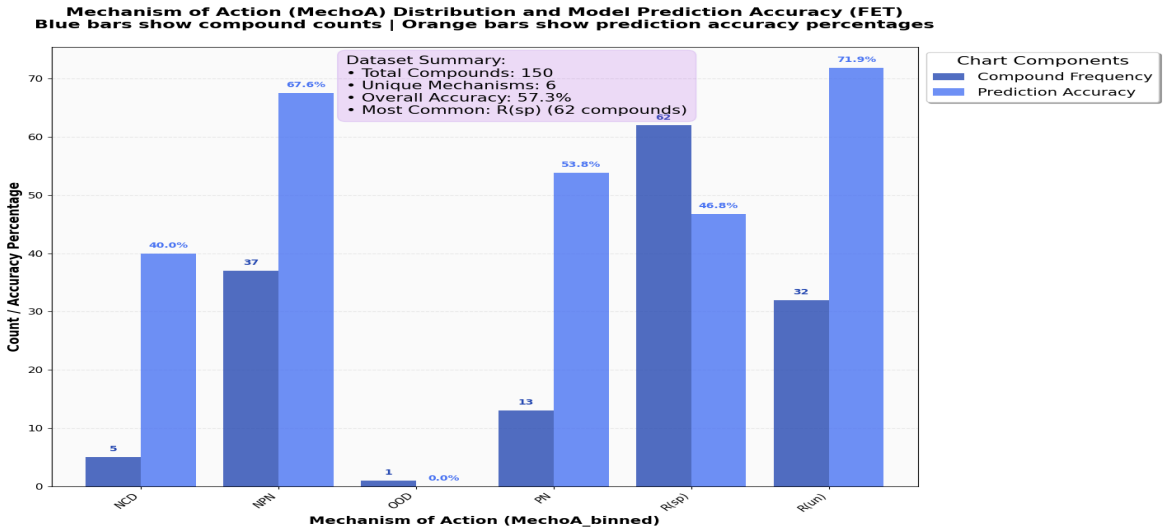
Based on our results, we hypothesized that the RT-gill test, as a simplified biological system, exhibits inherent limitations compared to whole-organism assessments. Specifically, this system may demonstrate reduced sensitivity to neurotoxic modes of action[15] and lacks the complex biological interactions present in intact organisms. Consequently, the RT-gill test may be more effective as a screening tool for identifying non-toxic substances rather than accurately predicting specific toxicity levels, particularly for chemicals operating through complex modes of action or intricate adverse outcome pathways.

This hypothesis can be backed using a 2022 study[16], which involved evaluating three testing approaches for whole effluent toxicity (WET) assessment: the traditional fish larvae acute toxicity test, the FET test using fathead minnow (Pimephales promelas), and the in vitro RTgill-W1 fish gill cell assay. Using 12 relevant chemicals, the study demonstrated that while FET showed significant correlation with fish larvae toxicity (LC50 values), the RTgill-W1 cell assay did not correlate well with whole organism responses. This results correlate our results and suggest some limitation due to the reduced biologicall complexity, and enforce that further type of test are needed to well predict the ecotoxicity.

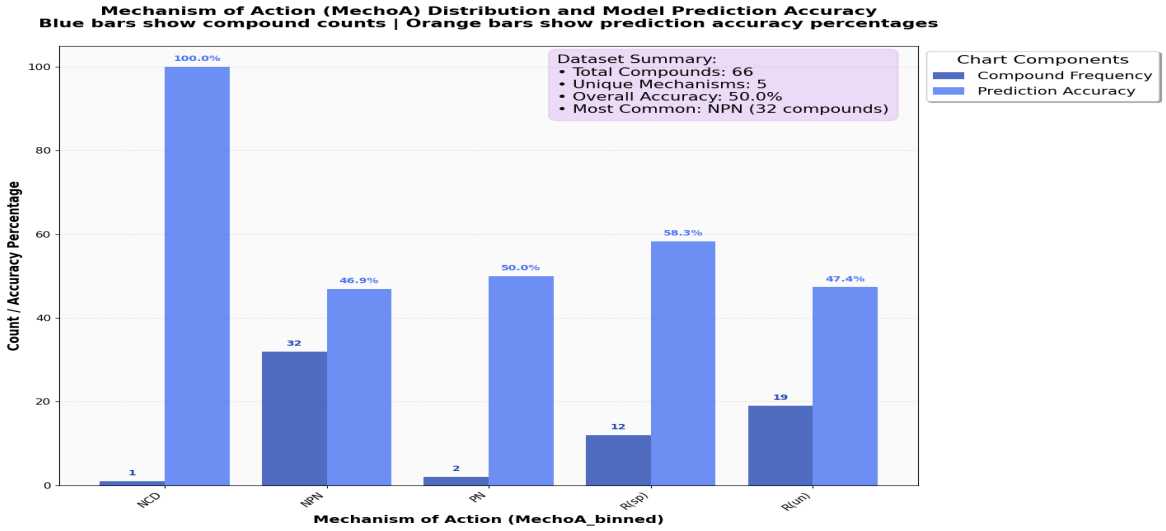
The following approach compared the performance, based on the mechanisms of action. **The idea behind this comparison was that for an untested chemical, on the basis of the predicted mechanism of action, one can choose the test to be performed which could give us a more accurate result.** The binning of the different mechanisms of action (hereon abbreviated as MechoA) into the broader classes was done with the reference paper nomenclature in mind:

- Polar Narcotic compounds - PN
- Non-polar narcotic Compounds - NPN
- Reactive (specific) compounds - R(sp)
- Reactive (unspecific) compounds - R(un)

- No chemical domain - NCD
- Out of Distribution - OOD



(a) *MechoA-wise performance for FET*



(b) *MechoA-wise performance for RTGill*

Figure 8: *MechoA-based comparison of RTGill and FET data*

Comparative Analysis

The reference data included the binned values of the mechanisms of action for each chemical which has been used for the analysis. FET test's highest percentage for correct prediction is for R(un) (71.9%), which also has a relatively high frequency (32) as compared to RTGill which has a lower frequency (19) and a lower accuracy (47.4%). Thus the data suggests that this group of compounds is better predicted by FET data.

FET also showed a relatively strong performance with a correct prediction percentage of 67.6% and high frequency (37) for NPN chemicals, indicating reliability for this mechanism compared to RTGill which despite having a comparable number of chemicals (32), did not impress with only 46.9% accuracy.

For R(sp), it is interesting to note that it was the highest occurring class for which FET data is available; however, it provided a sub-par performance with only 46.8% accuracy. Although the RTGill test provides

a higher accuracy for this mechanism, the result is based on only 12 compounds as compared to 62 in FET and hence cannot be compared. For the remaining classes PN and NCD, the number of compounds for RTGill data was too small (2 and 1, respectively) to have a fair comparison.

Within the constraints of the dataset used in this study, the Fish Embryo Toxicity (FET) test demonstrated greater versatility in predicting acute toxicity classes. Its predictive accuracy was particularly notable for chemicals with reactive-unspecific and non-polar narcotic mechanisms of action. The primary strength of the RTGill test, conversely, was its superior performance in correctly identifying non-toxic substances. However, the analysis did not yield sufficient evidence to suggest an improved performance for the RTGill test based on any specific mechanism of action. Therefore, a more extensive dataset is required to validate these preliminary observations and draw definitive conclusions regarding the optimal applications for each assay.

The implementation of comprehensive in-silico screening approaches for chemical evaluation, both toxicity prediction and mechanism of action analysis should be carried out for a new chemical. This should be integrated into decision-making frameworks that consider time-frame constraints and budget limitations (detailed in the Appendix) before proceeding with experimental testing. The predictive accuracy and reliability of these results could be substantially enhanced through the acquisition of larger, more diverse datasets, ultimately improving the cost-effectiveness and efficiency of chemical safety evaluation workflows.

5.5 Building Models

The next task involved building robust machine learning models with the goal of testing its performance on the reference data set to help further improve the accuracy using the defined approach.
The data were collected from the ECHA website and contained the following list of relevant features regarding the chemicals:

- CAS Number
- SMILES
- OPERATOR (It specifies whether the LC50 value in reality is higher ('>') or lower ('<') than the recorded value)
- LC50 value (mg/L)

There were other information, for example the year and source which can be used to check the relevance and trust regarding the experiments but that is beyond the scope of this project. We assume that all the data is relevant as of now and move on to perform some exploratory data analysis to help further clean and standardize the data as per our needs.

5.5.1 Molecular Fingerprints

Molecular fingerprints are mathematical representations of chemical structures that encode structural information into fixed-length binary vectors or variable-length hash sets. These representations enable rapid similarity searching, clustering, and machine learning applications in drug discovery, chemical database screening, and property prediction.

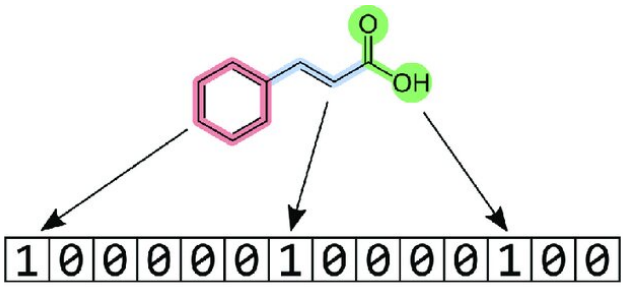


Figure 9: *Example of Molecular Fingerprint Generation[17]*

A molecular fingerprint $\mathbf{F}$ for a molecule M can be defined as:

$$\mathbf{F}(M) = \{f_1, f_2, \ldots, f_n\} \in \{0, 1\}^n$$

where each bit f_i represents the presence (1) or absence (0) of a specific structural feature.
The similarity between two molecules M_1 and M_2 is typically computed using the **Tanimoto coefficient** also known as the **Jaccard index**[18]:

$$T(M_1, M_2) = \frac{|\mathbf{F}(M_1) \cap \mathbf{F}(M_2)|}{|\mathbf{F}(M_1) \cup \mathbf{F}(M_2)|} = \frac{N_{11}}{N_{11} + N_{10} + N_{01}}$$

where N_{11} is the number of bits set to 1 in both fingerprints, N_{10} and N_{01} are the numbers of bits differing between the fingerprints.
To evaluate the best strategy for utilising molecular fingerprints, we compared two principled approaches: one that retains chemically informed structure (unfolded fingerprints), and one that evaluates purely statistical groupings (folded + dimensionality reduction). This comparison framework allows us to contrast the chemoinformatically informed versus the lack of domain knowledge and the more data-driven statistical representations. For more information and a detailed scheme regarding their use, check section 8.3.2.

5.5.2 Approach I - OECD Toolbox Data

Data Extraction and Pre-processing

Our approach deviated from typical structure-only methodologies by incorporating physical, chemical, biological, and structural data from the OECD QSAR toolbox, covering over 40,000 chemicals with additional calculations performed using EPI Suite for missing data. The data underwent comprehensive cleaning procedures including correlation-based feature removal using a greedy priority-based algorithm (preserving ”hub features” with highest correlation counts), removal of chemically inconsistent entries (LC50 values exceeding water solubility), and elimination of ambiguous operator-flagged data points. Molecular preprocessing involved desalting and stereochemistry removal to ensure standardized chemical representations. Both folded (Morgan + CACTVS) and unfolded Morgan fingerprints were generated and processed, with folded fingerprints undergoing dimensionality reduction via `TruncatedSVD` to 20 components (capturing 55% explained variance) to address curse of dimensionality concerns. The final dataset integrated toxicity data from ECHA with structural and physicochemical features, forming a comprehensive feature space for model training. Explicit details regarding all cleaning and pre-processing procedures, including algorithmic implementations and parameter justifications, are provided in Appendix Section 8.2.

Exploratory Data Analysis

Various different infographic plots were made to see the distribution of the data. Many of the features had discrete values (number of nitro groups, number of rings) and some of them had skewed distributions (Exp. Melting Point °C, Phi —Mackay Model). Basic statistics using the `describe()` function were checked and depending on the frequency of NaN values, if they formed the majority of the values the feature itself was removed, or we performed imputation by taking the geometric mean of the column values.

From the graph for the number of classes present in the training data, we can see that there exists an imbalance in the data that needs to be looked into.

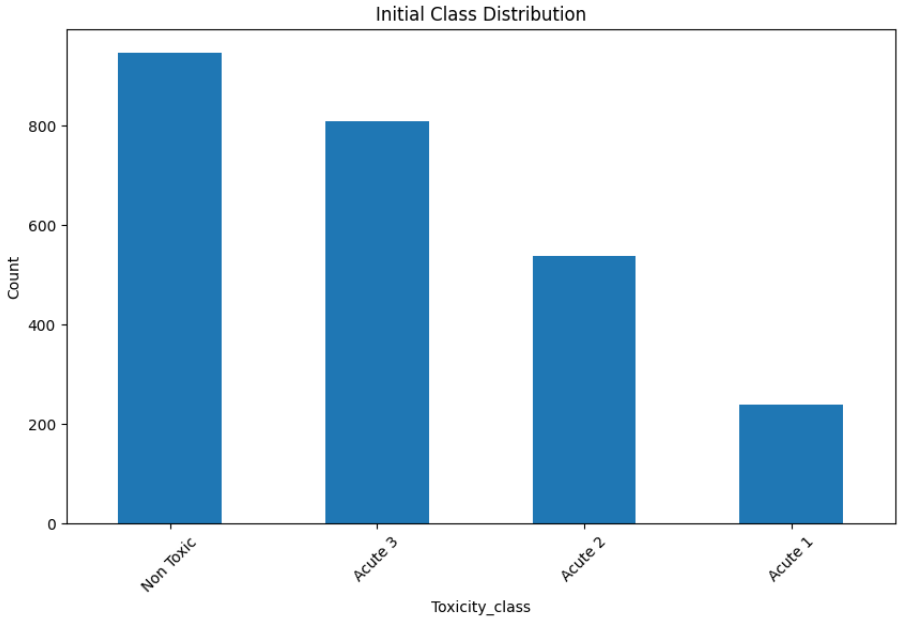


Figure 10: *Class-wise count of chemicals in training data*

This shows that there exists a major disparity between the majority class ('Non Toxic') and the minority class ('Acute 1'). It is to be remembered that we would rather have false positives, especially in the case of 'Acute 1' chemicals, than false negatives. Hence, it is necessary to address this issue using balancing techniques. In this regard, we decided to go for the **random oversampling method**. As we are dealing with biological and chemical data and their structures, methods like SMOTE (Synthetic Minority Over-sampling Technique) which may be effective in other use cases is not viable here, since it may interpolate and create impossible structural and biological data which in reality have a lot of domain restrictions. Hence, sticking to simply replicating the data points to counteract the imbalance seemed like the best option. In the end, we were left with **3283 data points for training**.

Test Data

For the test data, since the motivation for the project is to get our model included in the defined approach, we decide to use the reference data that were utilised earlier to compare and create the defined approach. However, in the case of the state-of-the-art models, since they are not open source we could check whether any of the chemicals in our reference data is already present in the training data or not. However, in our case, we can and hence did consider only those chemicals which are unknown to the model.

Model Training

For the training, we decided to implement various different models and decide the best based on the f1 score of the individual classes and the accuracy. Most of the models had their own packages which were imported individually while some of them were included in the `scikit-learn` package, like `Random Forest`.

Hyperparameter tuning was conducted using random search, implemented through Google Cloud Platform's Cloud Run Jobs. Training scripts were containerized using `docker` images to ensure consistent computational environments and reproducible results across multiple optimization runs. Each `docker` container included all necessary dependencies, pre-processing pipelines, and evaluation metrics to ensure isolated and reproducible training environments.

Hyperparameter search spaces were defined on the basis of algorithm-specific best practices. The optimization process focused on key parameters including learning rates, regularization terms, tree-based parameters (depth, number of estimators), and algorithm-specific settings such as `LightGBM`'s boosting parameters.

5.5.3 Approach II - Molecular Descriptor Data

Data Cleaning

In the second approach, instead of extracting the features from the OECD toolbox, we decided to use the molecular descriptors that are available for use on rdkit. The idea was to extract every available descriptor for all chemicals. This would then be subjected to cleaning on the basis of molecular weight, since any chemical having weight higher than 1000 Da could not permeate into the cells. Cleaning also took place based on the 'MolLogP' descriptor from the which contained the regressed log K_{ow} values for the chemicals. This coefficient is a critical measure of a chemical's lipophilicity, indicating its preference for a fatty environment (octanol) versus an aqueous one (water). A high log K_{ow} value signifies greater **hydrophobicity** (water-fearing), while a low value indicates **hydrophilicity** (water-loving). It is highly complex for this metric to be measured above a value of 6 experimentally and negative values indicate extreme hydrophilicity making the data inconsistent. Hence all values exceeding 6 and below 0 were removed.

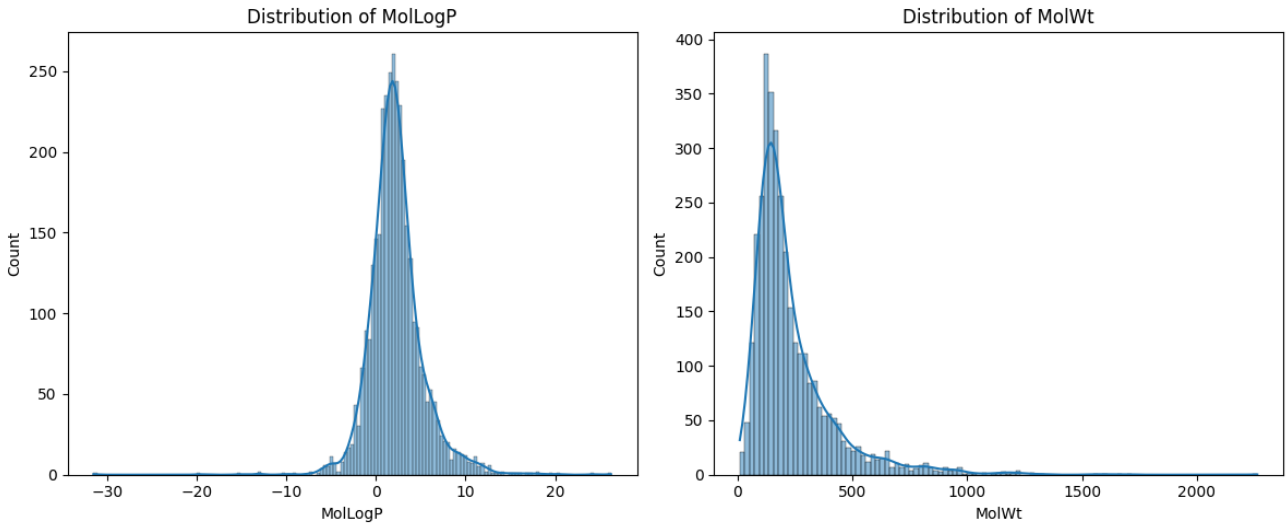


Figure 11: *Distribution of Log P and molecular weight in the data*

The rest of the steps in terms of cleaning the data with the help of the correlation matrix and also the

cleaning of the fingerprint data remain the exact same, since there are no changes in this regard. The only changes are in the chemicals themselves, since some of the chemicals which may have been left out earlier during the cleaning process may still remain and vice versa. However as the fingerprints were generated for all the chemicals in the data, the same could be used in this approach. This is because there is no water solubility descriptor and no cleaning took place in this approach using this aspect.

This was again followed by random oversampling of the data.The **test data** remains the same as well; however, since there were changes in the chemicals, again, the reference data were removed of all the chemicals which were in the training data.

There were no further changes. We proceeded with random oversampling again to handle the imbalance and experimented with both types of fingerprints to compare the two approaches. In the end, we had a total of **3496 data points available for training**. The model training process remained the same, with the same choice of hyperparameters and models.

5.5.4 Cluster Formation

We decided to cluster the data based on the two methods of fingerprint generation. The idea of clustering came as an attempt to try and improve the general model's performance. As specific smaller-sized models trained on chemicals of the same type may outperform a model trained on the whole data.

For clustering using the OECD toolbox data and the **folded Morgan fingerprints + CACTVS fingerprints**, there was no clarity when we used only the fingerprints (PC1 and PC2). We could not use metrics like the Tanimoto similarity as the concept of 'similarity' between bit-vectors did not remain after dimensionality reduction. Hence, we decided to map the data points using unsupervised clustering algorithms and in addition to the first 2 fingerprint components (containing the most amount of explained variance) used an additional component, the most important feature (which will be spoken of more in Section 4) for toxicity prediction of the general model. The most important feature for the general model was 'Water Solubility (fragments)', which calculates the water solubility of every fragment of the molecule followed by summing them up.

The option which gave us the best results for cluster formation was HDBSCAN. We reached this conclusion by testing two other algorithms, K-Means clustering and DBSCAN. Both of the methods and their respective optimization techniques that were undertaken have been detailed in the Appendix (section 8.5). Unlike DBSCAN's single-density assumption, HDBSCAN builds a hierarchy of clusters at different density levels, automatically selecting the most stable clusters. Also, DBSCAN requires fine-tuning both `eps` and `min_samples`, HDBSCAN primarily needs `min_cluster_size` which is the minimum number of points required to form a valid cluster in the final clustering solution, making it less prone to parameter sensitivity. For the search space of `min_cluster_size`, we vary between 1% and 10% of the total data size. `min_samples` ranges between 1 and the maximum value of `min_cluster_size`. We define a validity score method, which gives the silhouette score the highest (70%) weight, followed by cluster persistence which is HDBSCAN's internal stability metric and noise ratio (10%) which penalises solutions with excessive noise ($>$80%). We decided to take advantage of HDBSCAN's internal tools along with giving the silhouette score the highest weight, in line with the optimisation technique used for the previous two methods. The division of the weights was decided after multiple runs with different values keeping the hierarchy of the weights intact which did not lead to better clusters. The following solution providing the highest validation score provides us with the best clusters.

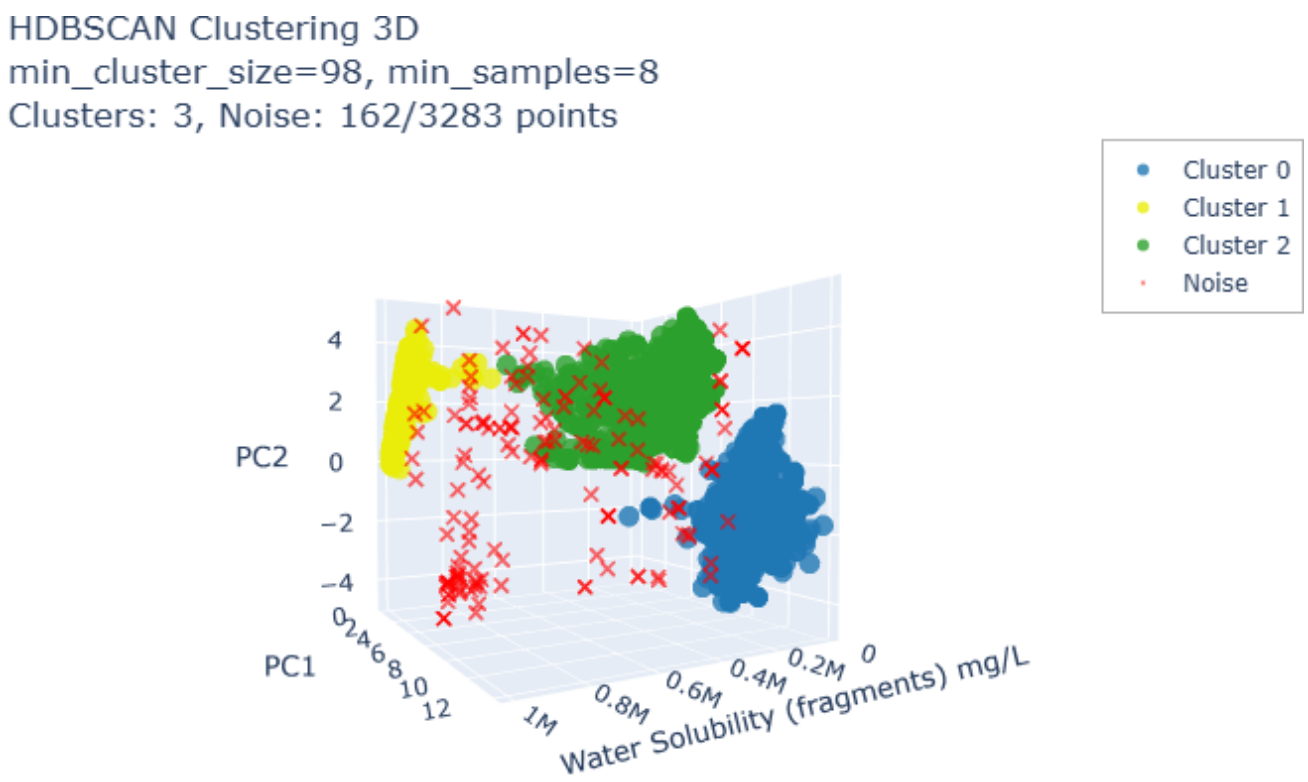


Figure 12: *Clustering with HDBSCAN*

The differences between DBSCAN and HDBSCAN include the observed cluster 1 (see section 8.5) in DBSCAN now broken down into two different clusters and also cluster 2 in DBSCAN which consisted of very few points is now classified as noise.

The model created using the combination of the molecular descriptor data and folded Morgan + CACTVS fingerprints provided us with sub-optimal performance and hence no clustering on the basis of this combination was produced. This has been clarified in section 5.

In the case of using the **unfolded Morgan fingerprints**, the similarity matrix was calculated using the Tanimoto similarity of each possible pair of chemicals. This was then translated into the distance matrix (1 - similarity matrix). The distance matrix was then compressed using UMAP before using HDBSCAN to visualise the clusters, to continue using the best clustering method. However, one is free to try various different clustering methods if required.

UMAP was chosen to compress the distance matrix, as it maintains both local neighborhoods and global topology and is the gold standard for dimensionality reduction currently.

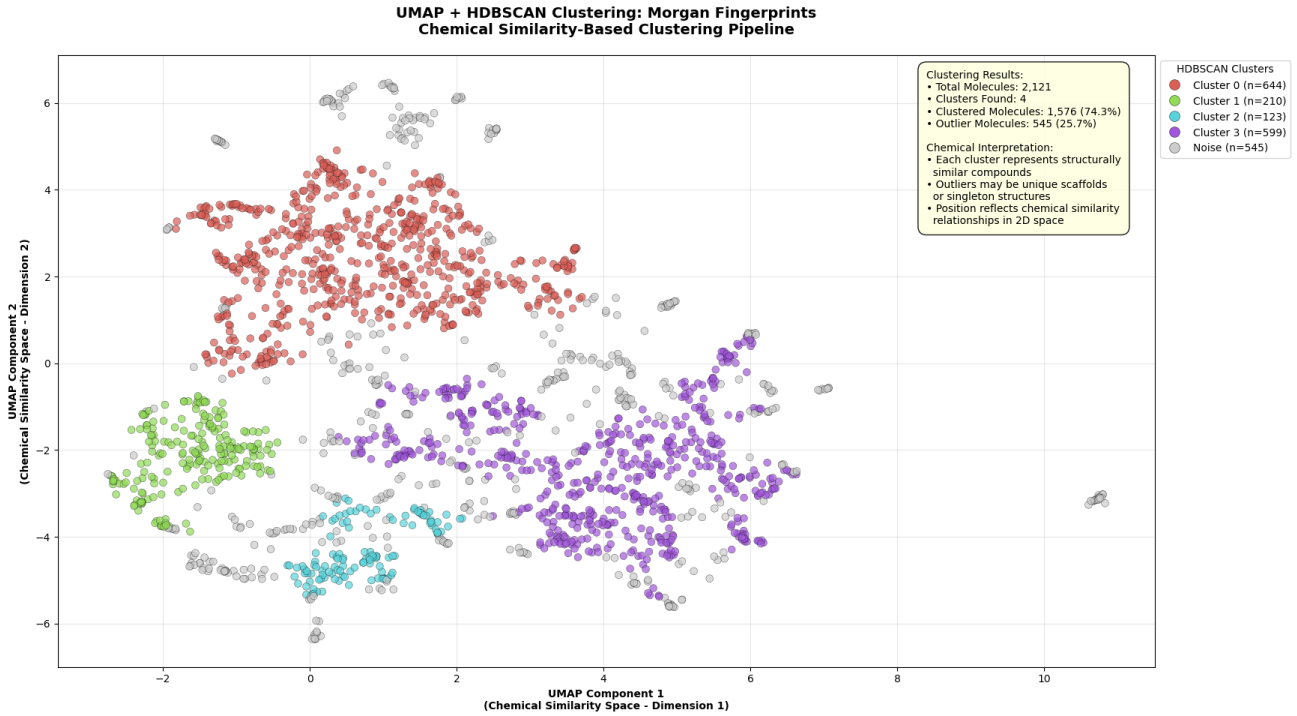


Figure 13: *Clustering for Unfolded Morgan Fingerprints*

Here we can observe that many small sub-clusters comprise the three main clusters making them quite sparse and no clear pattern can be observed.

Cluster based models

The motivation for cluster specific models stems from the idea of the creation of an applicability domain for our models. Also, the idea to have smaller sized models for specific types of chemicals can help us identify the toxicity class of new chemicals that may belong to a particular cluster better. Hence we took the best performing general model and tried to create models for the clusters that are formed for that particular data combination, to see if there is any improvement in prediction performance. Although the number of chemicals that can be used for testing for a particular cluster was less, we compared the predictions made for those chemicals by the cluster-specific model and the general model to understand the difference.

6 Results & Discussion

After the raw data was cleaned and structured, we ran a baseline model to understand the minimum accuracy performance. Thus, a logistic regression model without any hyper-parameter tuning was applied. This was done for the combination of both types of data, the OECD toolbox, and molecular descriptor data.

Class	Precision	Recall	F1-Score	Support
Acute 1	0.47	0.77	0.59	53
Acute 2	0.73	0.40	0.52	75
Acute 3	0.45	0.30	0.36	50
Non Toxic	0.39	0.67	0.49	24
Overall Accuracy	**0.50**			**202**

Table 4: *Logistic Regression Model Performance - OECD toolbox with folded Morgan + CACTVS fingerprints*

Class	Precision	Recall	F1-Score	Support
Acute 1	0.34	0.14	0.20	70
Acute 2	0.34	0.13	0.19	83
Acute 3	0.19	0.13	0.15	54
Non Toxic	0.06	0.29	0.10	28
Overall Accuracy	**0.15**			**235**

Table 5: *Logistic Regression Model Performance - OECD toolbox with unfolded Morgan fingerprints*

Class	Precision	Recall	F1-Score	Support
Acute 1	0.66	0.42	0.51	60
Acute 2	0.70	0.28	0.40	68
Acute 3	0.31	0.59	0.40	46
Non Toxic	0.29	0.60	0.39	20
Overall Accuracy	**0.43**			**194**

Table 6: *Logistic Regression Model Performance - Molecular descriptors with folded Morgan + CACTVS fingerprints*

Class	Precision	Recall	F1-Score	Support
Acute 1	0.63	0.28	0.39	60
Acute 2	0.38	0.19	0.25	68
Acute 3	0.23	0.11	0.15	46
Non Toxic	0.09	0.50	0.15	20
Overall Accuracy	**0.23**			**194**

Table 7: *Logistic Regression Model Performance - Molecular descriptors with unfolded Morgan fingerprints*

From the classification reports of the models it is quite clear when the features are from the OECD toolbox and are combined with the folded version of the Morgan fingerprints along with the CACTVS fingerprints, the predictivity of the baseline model (F1-score) seems to be the highest for every class except for Acute 3 which is 0.36. However it has 50 support points which is higher than for the rest of the approaches. Notice that the number of chemicals used for testing the model varies with each approach. This is because of the cleaning of the test data based on the training data that was noted earlier as well as omission of the chemicals from the test set which do not have the required features available.

Next, we present the results obtained using `catboost` and `random forest` models. Tree-based models succeeded because they're perfectly suited to the characteristics of molecular fingerprint and chemical features data: high-dimensional, binary, sparse features representing discrete molecular substructures and properties. The combination of training efficiency, interpretability, robustness to overfitting made

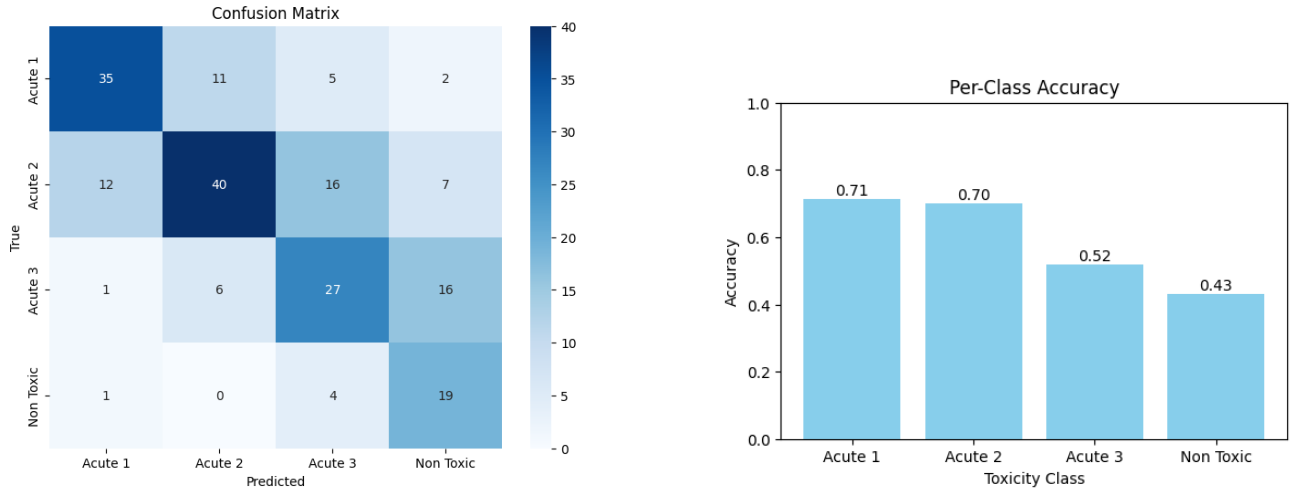
them the optimal choice for our ecotoxicity prediction task. Given the amount of data that we have, it becomes difficult for neural network based models to generalize, and hence our model using graph neural networks failed to compete.

We tested different combinations of data using a fixed random seed (41) for consistency and reproducibility. For each data combination, we tried multiple machine learning models and selected only the best-performing one for further optimization. The results shown represent these top models after hyperparameter tuning. To ensure fair comparison, we used the same starting parameters when evaluating which model to select for tuning, regardless of the data combination or random seed used.

6.1 OECD toolbox with folded Morgan + CACTVS fingerprints

Class	Precision	Recall	F1-Score	Support
Acute 1	0.71	0.66	0.69	53
Acute 2	0.70	0.53	0.61	75
Acute 3	0.52	0.54	0.53	50
Non Toxic	0.43	0.79	0.56	24
Overall Accuracy	**0.60**			**202**

Table 8: *Classification Report - Random Forest Classifier*



(a) *Confusion Matrix* ***(b)*** *Per class accuracy*

Figure 14: *Performance results*

The model demonstrates a good performance for the 'Acute 1' class, achieving the highest precision (0.71) and the second best recall (0.66). This is particularly noteworthy given that 'Acute 1' was a minority class in the original training data. 'Acute 2', being the most represented class in the test set (n=75), shows good precision (0.70) but moderate recall (0.53). While the model makes reliable predictions for this class, it misses nearly half of the true 'Acute 2' cases.

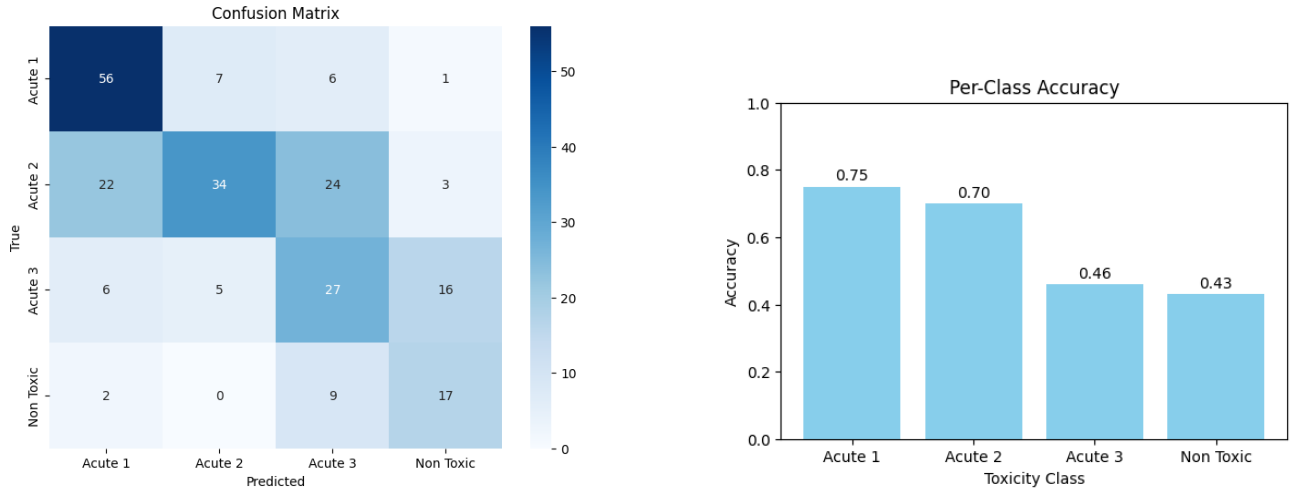
The 'Non Toxic' class presents an interesting pattern with the highest recall (0.79) but the lowest precision (0.43). The model successfully identifies most non-toxic compounds but generates many false positives. Given the smallest test sample size (n=24), this assessment should be interpreted cautiously, though the high recall suggests the model may be over-predicting this class.

'Acute 3' continues to show the characteristic challenge in ecotoxicity prediction with moderate precision (0.52) and recall (0.54). This aligns with the general observation that ecotoxicity prediction models struggle with 'Acute 3' classifications due to the complex and subtle chemical patterns exhibited by compounds in this category.[4]

6.2 OECD toolbox with unfolded Morgan fingerprints

Class	Precision	Recall	F1-Score	Support
Acute 1	0.75	0.71	0.73	70
Acute 2	0.70	0.53	0.60	83
Acute 3	0.46	0.52	0.49	54
Non Toxic	0.43	0.68	0.53	28
Overall Accuracy	**0.60**			**235**

Table 9: *Classification Report - Catboost*



(a) *Confusion Matrix* ***(b)*** *Per class accuracy*

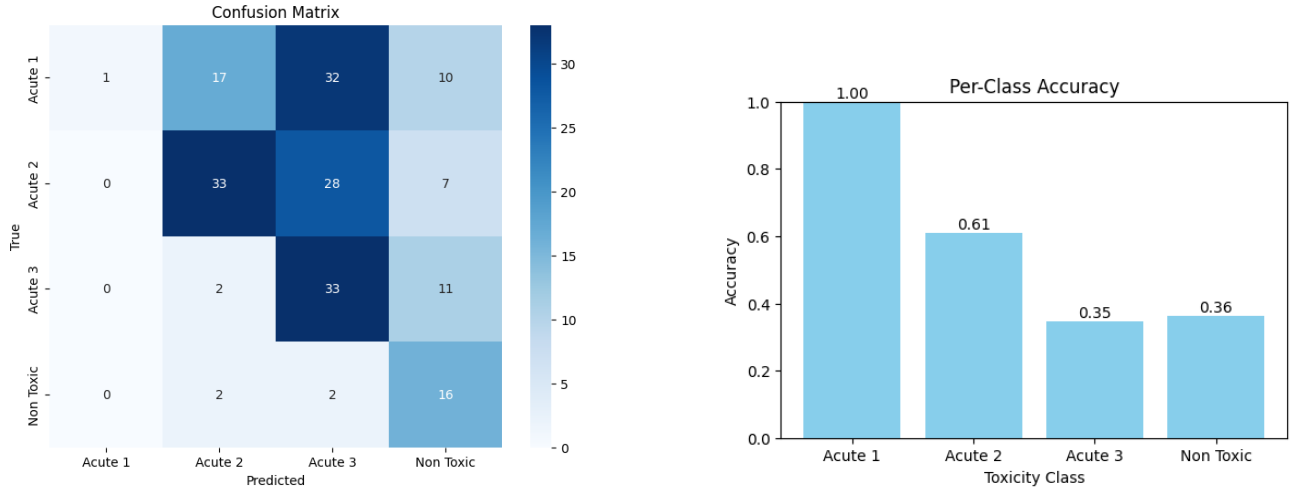
Figure 15: *Performance results*

We see a very similar result using the unfolded version of the Morgan fingerprints as we did with the previous result. The overall accuracy remains the same but with a relatively higher number of test chemicals. The trend of a higher 'Acute 1' and 'Acute 2' accuracy and lower accuracy for the 'Non Toxic' and 'Acute 3' classes remain. However, it is important to note that the 'Acute 3' class has a lower F1-score than before with a lower precision and recall score. The support for 'Acute 1' increased from 53 to 70 with a relatively higher precision and recall indicating a more robust performance.

6.3 Molecular descriptors with folded Morgan + CACTVS fingerprints

Class	Precision	Recall	F1-Score	Support
Acute 1	1.00	0.02	0.03	60
Acute 2	0.61	0.49	0.54	68
Acute 3	0.35	0.72	0.47	46
Non Toxic	0.36	0.80	0.50	20
Overall Accuracy	**0.43**			**194**

Table 10: *Classification Report - Catboost*



(a) *Confusion Matrix* ***(b)*** *Per class accuracy*

Figure 16: *Performance results*

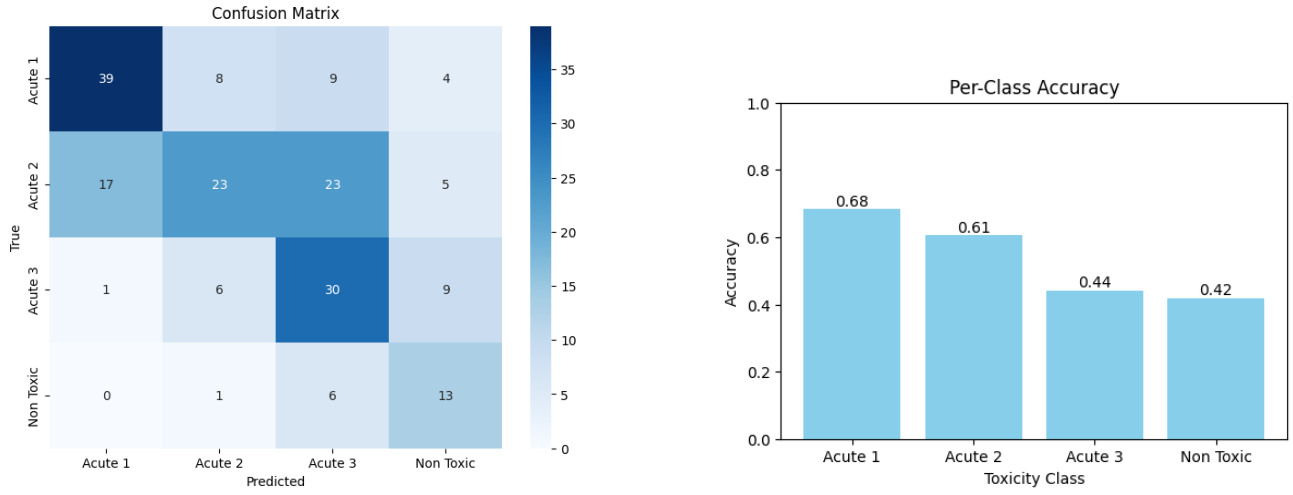
The model shows mixed performance across ecotoxicity classes. 'Acute 1' achieved perfect precision (1.00), meaning when it predicts this class, it's always correct. However, it has extremely poor recall (0.02), indicating it identifies only 2% of actual 'Acute 1' cases

'Acute 3' and 'Non Toxic' classes demonstrate the exact opposite pattern with high recall and low precision values, hence correctly identifying most true cases but generating many false positives.

6.4 Molecular descriptors with unfolded Morgan fingerprints

Class	Precision	Recall	F1-Score	Support
Acute 1	0.68	0.65	0.67	60
Acute 2	0.61	0.34	0.43	68
Acute 3	0.44	0.65	0.53	46
Non Toxic	0.42	0.65	0.51	20
Overall Accuracy	**0.54**			**194**

Table 11: *Classification Report - Catboost*



(a) *Confusion Matrix* ***(b)*** *Per class accuracy*

Figure 17: *Performance results*

We can notice the same trend in terms of accuracy repeating in this combination of the data as well. However there is no improvement in the overall accuracy of the predictions as compared to the previous models.

Based on the overall accuracy and the per class performance, the combination of OECD data along with both the approaches in terms of the fingerprints gave the same overall accuracy, while in terms of the F1-score the first combination does better in 3 out of the 4 classes. It is to be noted that when the unfolded version of the Morgan fingerprints were being used, the total time taken to tune a model was significantly more than when we used the dimensionally reduced version of the combination of the folded Morgan and CACTVS fingerprints.

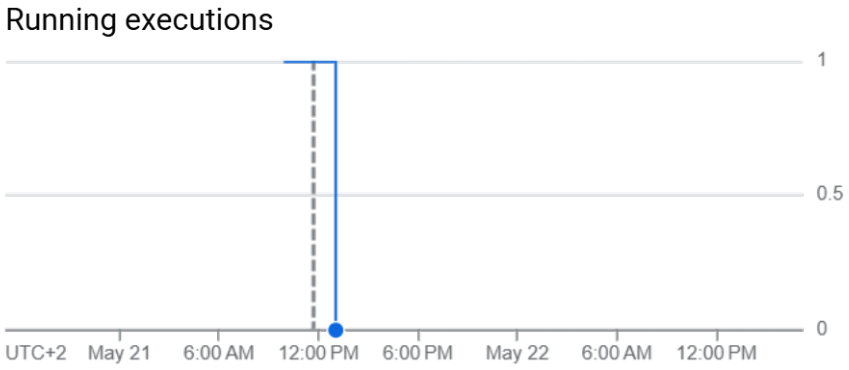


Figure 18: *Example of execution time for folded fingerprint approach*

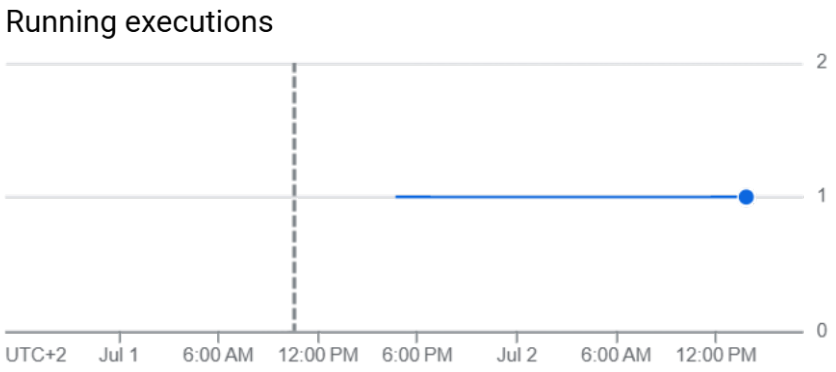


Figure 19: *Example of execution time for unfolded fingerprint approach*

For the folded version it took somewhere between 3.5 hours and 4 hours for the job to be completed, while for the unfolded version it took close to 13 hours. This major difference would suggest that the best model is that of the combination of OECD toolbox data and the folded Morgan and CACTVS fingerprints. Not only is is providing us with the best performance in terms of predictivity but also is almost 2.5 times faster then its second best counterpart. Therefore, we further defined models based on the clusters that were found when using this version of the features for our data.

6.5 Cluster specific models

The last investigation of the project was to try and create models that would cater to the test chemicals belonging to the specific clusters that were defined using the fingerprints and 'Water Solubility' feature in our first approach. The objective was to create models that would be better at predicting the ecotoxicity of these particular chemicals than the general model that was defined.

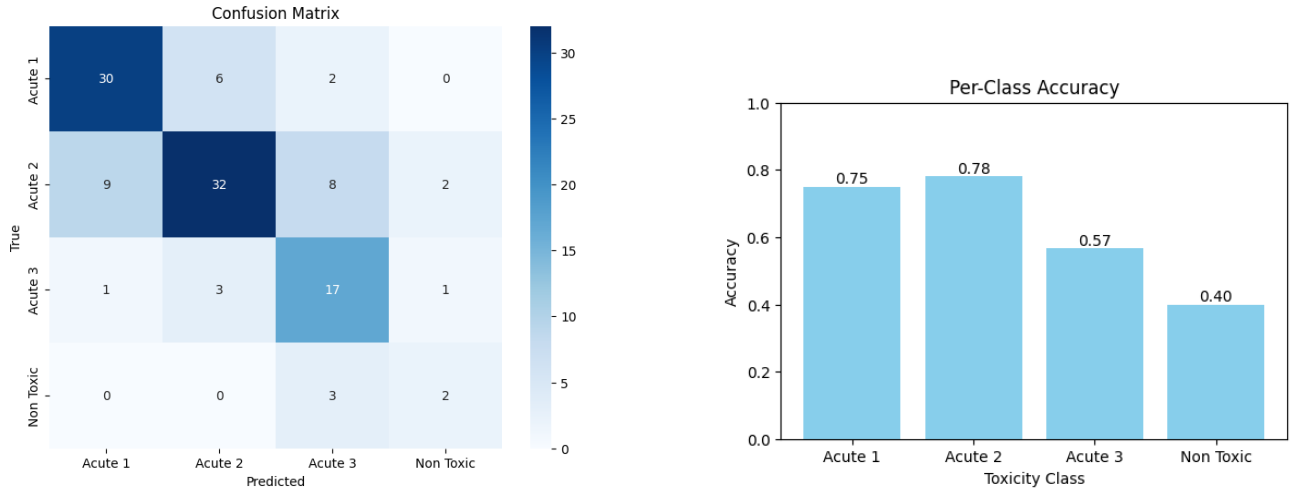
Based on the 3 clusters that can be seen in figure 12, the goal was to have 3 different models that would be better at predicting the unseen chemicals that would be assigned to the clusters. This would also help us define an applicability domain for the model overall, as any chemical not belonging to the 3 clusters would be classified as out of domain for our general model.

When we look into the number of training and testing data points per cluster, we can find that for cluster 0, we have 1169 training points and 116 testing points. For cluster 1, we have 401 training points and only 7 test chemicals, while for cluster 2 we have 1567 training data points and 77 testing chemicals to create our models. Hence, we proceeded to work on building models for clusters 0 and 2 only, since we did not have enough chemicals for us to test the model, if built, using the cluster 1 data.

6.5.1 Cluster 0

Class	Precision	Recall	F1-Score	Support
Acute 1	0.75	0.79	0.77	38
Acute 2	0.78	0.63	0.70	51
Acute 3	0.57	0.77	0.65	22
Non Toxic	0.40	0.40	0.40	5
Overall Accuracy	**0.70**			**116**

Table 12: *Classification Report - Random forest classifier*



(a) *Confusion Matrix* ***(b)*** *Per class accuracy*

Figure 20: *Performance results*

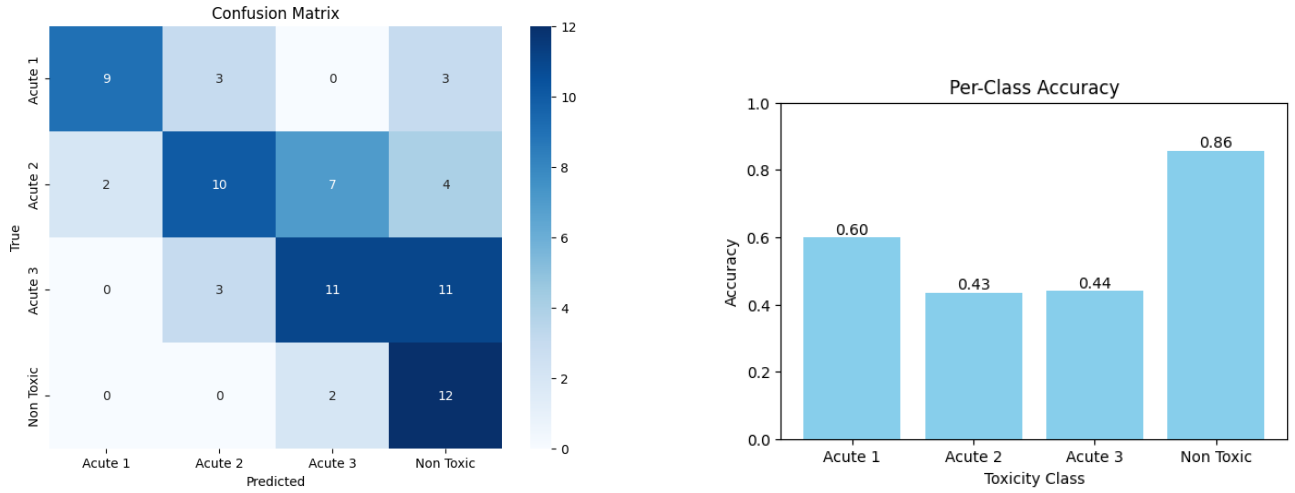
From figure 20 it is visible that there is an improvement in performance, however, the accuracies are based on a reduced test set and hence it is difficult to know based on the performance alone whether there was any improvement or difference in ecotoxicity predictions between the general and cluster-based model for the common chemicals. We performed further analysis and found that out of the 116 chemicals in the test set, 65 of them were differently predicted by the cluster-based model as compared to the general model. In total, the general model made 71 errors in classification as compared to 35 errors by the cluster-specific model, thus reducing mis-classifications to half for this set of chemicals. The number of instances where the cluster-based model is incorrect while the general model is correct is 10, as compared to 46 for the opposite case.

This clearly shows that, in the case of chemicals belonging to cluster 0, the cluster-based model outperforms the general model and is the one to be used.

6.5.2 Cluster 2

Class	Precision	Recall	F1-Score	Support
Acute 1	0.82	0.60	0.69	15
Acute 2	0.62	0.43	0.51	23
Acute 3	0.55	0.44	0.49	25
Non Toxic	0.40	0.86	0.55	14
Overall Accuracy	**0.55**			**77**

Table 13: *Classification Report - Random forest classifier*



(a) *Confusion Matrix* ***(b)*** *Per class accuracy*

Figure 21: *Performance results*

Again, to understand the differences in predictions, we looked into common chemicals to find the same trend as the previous case, with the cluster-based model outperforming the general model. There were disagreements in classification for 62 test chemicals out of 77 between the two models. The general model being incorrect 54 times out of 77, and the cluster-specific model being incorrect 35 times. The general model provided correct predictions for 14 chemicals for which the cluster-specific model was incorrect, while the cluster-specific model correctly predicted the ecotoxicity class for 33 chemicals for which the general model was incorrect.

Overall, we can see that the cluster-specific models outperformed the general model whenever we had enough data to train and test them.

7 Conclusion & Future Work

This internship project successfully navigated the complex landscape of non-animal alternatives for acute fish ecotoxicity testing by pursuing two primary objectives: the evaluation of a ”Defined Approach” (DA) integrating in vitro and in silico models, and the development of novel, standalone machine learning models.

A significant part of the study was dedicated to a comparative analysis of the RTGill-W1 and Fish Embryo ecotoxicity (FET) tests. The findings, while constrained by a limited set of 21 common chemicals, were insightful. Preliminary results suggest that the FET test may offer superior accuracy for predicting ecotoxicity in specific classes (Acute 2 and 3) and for chemicals with certain mechanisms of action. Conversely, the RTGill test showed greater efficacy in correctly identifying non-toxic substances. This suggests a potential complementary role for the two assays rather than a direct replacement. However, the analysis highlighted the critical need for a larger dataset, estimated at a minimum of 385 chemicals (section 8.1), to draw statistically significant conclusions and definitively guide future testing strategies.

Then we focused on building robust machine learning classifiers with the goal of getting our model included in the OECD toolbox for future use in a defined approach framework. The research concluded that a Random Forest model, utilizing a combination of OECD toolbox data with folded Morgan and CACTVS fingerprints, emerged as the most effective general model. It achieved a respectable overall accuracy of 60% and demonstrated a clear advantage in computational efficiency, being approximately 2.5 times faster to train than models using unfolded fingerprints without sacrificing predictive power.

The most promising breakthrough of this project came from the development of cluster-specific models. By segmenting chemicals into distinct clusters using the HDBSCAN algorithm based on their structural and physicochemical properties, we were able to create specialized models with clearly defined applicability domains. For chemicals falling within these domains (e.g., Cluster 0), the specialized models dramatically outperformed the general model, reducing mis-classification errors by as much as 50%.

Building on the insights from this project, future work will explore two primary avenues for improving predictive accuracy. First, we will investigate the creation of a new general model by exploring a strategic **middle ground** for feature representation. This will involve using the combined folded Morgan and CACTVS fingerprints *without applying any dimensionality reduction*. This approach explores a feature space that, with its $\sim$1900 dimensions, is significantly more detailed than the 20 dimensions retained by `TruncatedSVD`. However, it is vastly more constrained and computationally manageable than the high-dimensional, variable-length space of unfolded fingerprints. Crucially, the number of features remains below the number of training data points. The objective is to test the hypothesis that the complete, uncompressed feature set contains subtle structural information that was lost during the `TruncatedSVD` step, potentially yielding a model with higher predictive accuracy despite the anticipated increase in computational cost.

Second, we propose a knowledge-based data segmentation strategy centered on the **octanol-water partition coefficient (log K_{ow})**. Acknowledging the experimental difficulty in measuring log K_{ow} values above 6 and the physical impossibility of values below 0, we will partition the dataset into two distinct subsets: chemicals with a log K_{ow} between 0 and 3, and those between 3 and 6. For the OECD toolbox data, this will involve converting relevant thresholds to the soil organic carbon-water partition coefficient (K_{oc}) using established empirical formulas (e.g., $\log K_{oc} \approx 0.7919 \cdot \log K_{ow} + 0.0784$[19]), while for the molecular descriptor data, the division will be directly implemented using the existing 'MolLogP' feature. By training and testing models on these separate chunks, we aim to determine if specialized models tailored to specific lipophilic profiles can yield superior predictive performance.

References

[1] OECD. Test no. 249: Fish cell line acute toxicity - the rtgill-w1 cell line assay. 2021.

[2] EFPIA. Putting animal welfare principles and 3rs into action. 2010.

[3] European Commission. Roadmap towards phasing out animal testing. 2024.

[4] D. S. Macmillan, P. Ambure, V. Aranda, Y. Bayona, V. Bonderovic, J. Dawick, N. Fabre, S. Fischer, G. Hodges, Á. Llobet-Mut, S. Loisel-Joubert, C. Rivetti, J. Roberts, K. Schirmer, E. Serrano-Candelas, B. Serrano Ramón, and R. A. Stackhouse. Addressing the challenges of acute toxicity hazard classification using a non-animal defined approach. *Environ Toxicol Chem.*, 2025.

[5] L. Zhou, D. Fan, W. Yin, W. Gu, Z. Wang, J. Liu, Y. Xu, L. Shi, M. Liu, and G. Ji. Comparison of seven in silico tools for evaluating of daphnia and fish acute toxicity: case study on chinese priority controlled chemicals and new chemicals. *BMC Bioinformatics, 22(1)*, 2021.

[6] F. J. Bauer, P. C. Thomas, S. Y. Fouchard, and S. J. Neunlist. A new classification algorithm based on mechanisms of action. *Computational Toxicology, 5, 8–15.*, 2017.

[7] T.E.H. Allen, J.M. Goodman, S. Gutsell, and P.J. Russell. Defining molecular initiating events in the 443 adverse outcome pathway framework for risk assessment. *Res. Toxicol. 27, 2100–444 2112.*, 2014.

[8] KREATiS. Kreatis - mechoapedia, 2025.

[9] KREATiS. isaferat mechanisms of toxic action profiler v1.1. 2018.

[10] OECD. Oecd qsar toolbox v4.7. 2024.

[11] OECD. Guidance document on the reporting of defined approaches to be used within integrated approaches to testing and assessment. *OECD Series on Testing and Assessment, No. 255, OECD Publishing, Paris*, 2017.

[12] Istituto di Ricerche Farmacologiche Mario Negri IRCSS. Fathead minnow lc50 model (knn/irfmn) - v. 1.1.1. 2022.

[13] Istituto di Ricerche Farmacologiche Mario Negri IRCSS. Fish acute toxicity read-across version 1.0.1. 2022.

[14] R. T. Wright, K. Fay, A. Kennedy, K. Mayo-Bean, K. Moran-Bruce, Office of Pollution Prevention, Toxics, W. Meylan, P. Ranslow, M. Lock, J. V. Nabholz, J. Von Runnen, L. M. Cassidy, J. Tunkel, Inc. U.S. Environmental Protection Agency, SRC, and LLC. Consortium for Environmental Risk Management. Operation manual for the ecological structure-activity relationship model (ecosar) class program estimating toxicity of industrial chemicals to aquatic organisms using the ecosar (ecological structure activity relationship) class program ms-windows version 2.2. 2022.

[15] C. Schür, M. Paparella, C. Faßbender, G. Stoddart, M. B. Jesi, and K. Schirmer. Daphnids can safeguard the use of alternative bioassays to the acute fish toxicity test: A focus on neurotoxicity. *Environmental Toxicology and Chemistry.*, 2025.

[16] J. Scott, R. Grewe, and M. Minghetti. Fish embryo acute toxicity testing and the rtgill-w1 cell line as in vitro models for whole-effluent toxicity (wet) testing: An in vitro/in vivo comparison of chemicals relevant for wet testing. *Environmental Toxicology and Chemistry*, 41(11):2721–2731, 2022.

[17] J. Menke, J. Massa, and O. Koch. Natural product scores and fingerprints extracted from artificial neural networks. *Computational and Structural Biotechnology Journal*, 9:4593–4602, 2021.

[18] J.D. Holliday, C.-Y. Hu, and P. Willett. Grouping of coefficients for the calculation of intermolecular similarity and dissimilarity using 2d fragment bitstrings. *Combinatorial Chemistry and High Throughput Screening.*, 5(2) pp.155166, 2002.

[19] US EPA. Part 5: Chemical-specific parameters.

[20] W.G. Cochran. *Sampling Techniques, 3rd Edition.* John Wiley & Sons, New York., 1977.

[21] United States Environmental Protection Agency. Epi suite, 2025.

[22] G. Anand, P. Koniusz, A. Kumar, L. A. Golding, M. J. Morgan, and P. Moghadam. Graph neural networks-enhanced relation prediction for ecotoxicology (grape). *Journal of Hazardous Materials*, 2024.

[23] Wikipedia contributors. Birthday problem. *In Wikipedia, The Free Encyclopedia.*, 2024.

8 Appendix

8.1 FET and RTGill sample estimations and costs

The condition that needs to be verified for us to use the Normal approximation method as a rule of thumb is $np \geq 10$ and $n(1-p) \geq 10$ for both the tests:

$$
\begin{aligned}
np_{\text{fet}} &= 21 \times \frac{13}{21} = 13 \\
np_{\text{rtgill}} &= 21 \times \frac{11}{21} = 11 \\
n(1-p_{\text{fet}}) &= 21 \times \left(1 - \frac{13}{21}\right) = 21 \times \frac{8}{21} = 8 \\
n(1-p_{\text{rtgill}}) &= 21 \times \left(1 - \frac{11}{21}\right) = 21 \times \frac{9}{21} = 9
\end{aligned}
$$

In the absence of reliable preliminary estimates due to the small sample size (n=21), we used $p = 0.5$ in the sample size calculation for FET data, following standard statistical practice (Cochran, 1977)[20]. Since RTGill data provides a p already close to 50% we did not find it necessary to alter the value. This approach maximizes the variance term $p(1-p)$, providing the most conservative sample size estimate that ensures adequate precision regardless of the true accuracy values of either testing method.

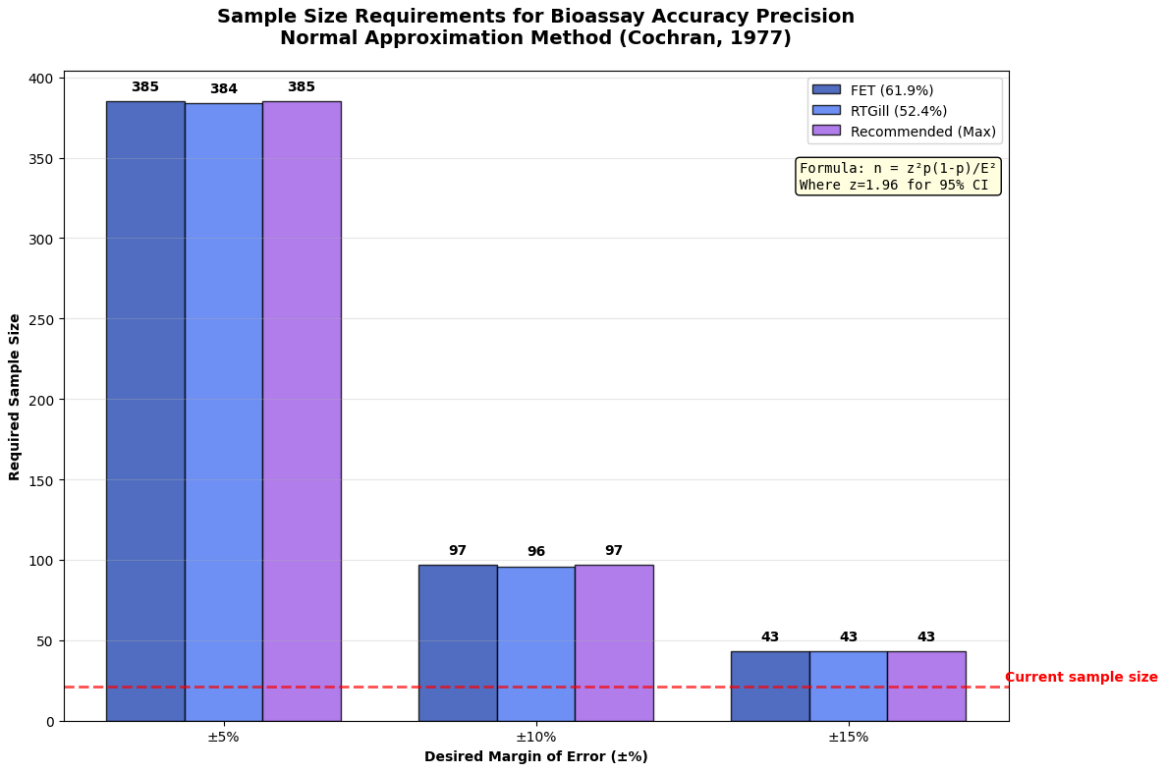


Figure 22: *Estimation of required sample size*

We can estimate that with **385 chemicals** having both FET and RTGill test data, we will have a **95% confidence interval with a 5% margin of error**.

After a discussion with an ecotoxicologist in our team who was responsible for handling and ordering of the tests, I got to know that in terms of costs, since there are approximately 7-8 labs across Europe

who perform the FET tests, depending on the chemical being tested, it can vary between 3,500 euros and 10,000 euros. While in the case of the RTGill test, it can vary between 6,000 and 7,000 euros. However, there exists only 1 lab in Europe based out of Switzerland, that performs the RTGill tests which was responsible for the ideation of the test itself. Hence, the waiting time to get back results can vary between 5-6 months, while for the FET tests, due to more labs having internalized the test takes relatively lesser time, around 3 months in general. This information gives us more context as to which test to perform based on the budget and time frame in which the results will be required.

8.2 Cleaning and pre-processing of OECD toolbox data

For the features of the model, the idea was to use the physical, chemical, biological, and structural data of the chemicals. For context, most state-of-the-art models depend on structure-based molecular descriptors and/or fingerprints for the predictions. Hence, this was a deviation from that direction since in this approach we chose to include the features from the OECD QSAR toolbox. For the features that are retrieved, there are more than 40,000 chemicals that have their information present in the database of the toolbox. However, for those that are not, the calculations are performed using the EPI Suite platform, using various regression models [21]. The initial screening of all relevant 2D properties was done using the domain knowledge of my supervisor. This was further followed by checking the correlation matrix and removal of the features that were highly correlated (threshold of 70 percent was fixed after multiple iterations). To determine which of the multiple features were to be kept, we decided to implement a greedy approach. We designed a Python function that rather than arbitrarily selecting which features to remove from correlated pairs, implemented a sophisticated priority-based removal strategy that processed features in descending order of their correlation count. The features with the highest correlations were treated as “hub features” and given priority to remain in the dataset, while their correlated partners were removed.

Algorithm: Greedy Priority-Based Correlation Removal

```
1. Create absolute correlation matrix from DataFrame
2. Extract upper triangle to avoid duplicate pairs
3. Identify features with correlations > threshold (0.7)
4. For each feature: count how many correlations it has
5. Sort features by correlation count (descending order)
6. Process features as "hubs" in priority order:
   - Keep the hub feature (highest correlation count)
   - Remove all its correlated partners
   - Skip if feature already marked for removal
7. Return DataFrame with highly correlated features removed
```

This effectively preserves potentially important central features that may capture key relationships in the data, while eliminating redundant variables that could cause multicollinearity issues.

One of the features extracted from the OECD toolbox was the water solubility of the chemicals. It is not logical for the LC50 value of the chemical to exceed the water solubility concentration, as it would mean that the lethal concentration exceeded the amount of the chemical that is soluble in water. Hence, of the 90,000+ rows, whenever the LC50 values exceeded the water solubility concentration, the row was removed.

Another feature was ‘OPERATOR’ which is available in the dataset itself, signifying if the actual LC50 exceeds or falls short of the value which is present in the data. For example, if the LC50 value of a chemical is 85 mg/L, and the operator is a ’>’, this means that the value is higher than noted. Hence, as we do not know the actual value, the data point becomes ambiguous in nature and was removed.

In the end, for every CAS Number with multiple values of LC50 available, we perform the geometric mean (similar to reference paper) to merge and hence classify the chemical using a simple Python function. Both the data, the one retrieved from the toolbox and the one extracted from ECHA are then merged based on the 'CAS Number' column.

Molecular pre-processing was performed to ensure consistent and standardized chemical representations prior to fingerprint generation. The `mol` object was first desalted using the `SaltRemover` class from the `rdkit` package, followed by removal of any stereochemistry using `RemoveStereochemistry()` from the `Chem` package. These standardization steps collectively ensure that fingerprint generation captures the core structural features relevant to the biological endpoint while minimizing noise from pharmaceutical formulation artifacts and stereochemical ambiguities.

Next using the `rdkit` and the `pubchempy` packages, the molecular fingerprints were generated that provided us with the structural data necessary about the chemicals in our data. Initially, we decided to use the folded version of the Morgan fingerprints along with the CACTVS fingerprints. This exact approach can also be found in the paper 'Graph neural networks-enhanced relation prediction for ecotoxicology (GRAPE)'[22] which emphasizes that the Morgan fingerprints capture local and global structural information using a radius-based algorithm, while the CACTVS fingerprints include predefined substructures and captures various chemical properties and functionalities, together, these fingerprints provide a comprehensive set of features for similarity-based comparisons among chemical molecules. For some of the chemicals, all the bits of the fingerprint were 0, and there was no use to have the structural data for those data points which form an essential part of the overall data for the models, and hence all such rows were removed. After cleaning the fingerprints, using the `sklearn` package we imported `TruncatedSVD` to reduce the dimensions of the fingerprints, which in our case was 1024 columns for the Morgan fingerprints and 881 for the CACTVS fingerprints.

As the total number of features, if kept as is, could lead to 'curse of dimensionality' issues (1900+ columns with < 4000 chemicals), dimension reduction was performed. It is also possible to keep folding the Morgan fingerprints to a desired value; however, doing so for the CACTVS fingerprints is not a possibility. The reduction of dimensions is a choice that we took as we had already lost the interpretability factor using the folded fingerprints, and one can also experiment without doing so. Two methods were experimented with for the dimensionality reduction, `SparsePCA` and `TruncatedSVD`, both known to handle sparse data (like fingerprints) well, out of which we decided to continue with `TruncatedSVD`. Further explanations and details regarding their differences have been mentioned in section 8.4. After multiple iterations on the basis of explained variance of the components (using `TruncatedSVD`), we decided to keep the first 20 components of the fingerprint data, having a cumulative explained variance of approximately 55 percent. With the 20th component contributing merely 0.71% explained variance, further components would yield minimal information gain relative to the added computational complexity justifying the 20-component cutoff. This was then merged with the earlier merged data to form our final dataframe which will be used for training.

In parallel we decided to test the unfolded version of the fingerprints to see if there is any changes in performance. All the steps for data extraction and generation remain the same except for the fingerprint generation function which instead of using `AllChem.GetMorganFingerprintAsBitVect(mol, radius=4, nBits=1024)`, used `AllChem.GetMorganGenerator(radius=4)`. We also do not perform any dimensionality reduction ignoring the 'curse of dimensionality' scare to keep the essence of using the unfolded version intact.

After this was done, all the identity features like the 'CAS Number' and 'SMILES' were removed as they were of no use to the models. Also, the 'Experimental_toxicity_mg_L' feature was removed since in the real world context, we will not have any data for this feature for an unknown chemical.

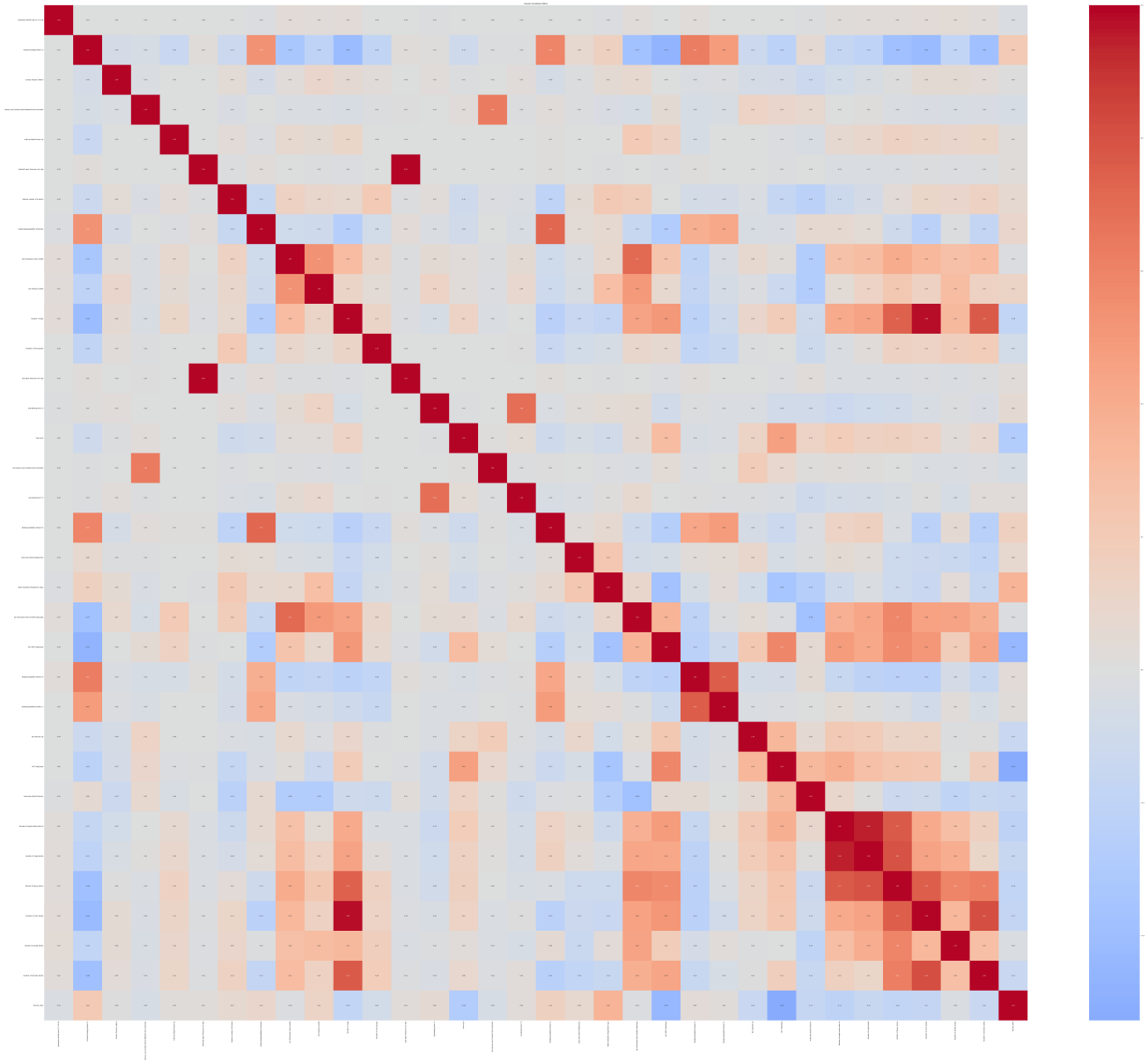


Figure 23: *The initial correlation matrix for the OECD toolbox data*

8.3 Types of molecular fingerprints and other features

8.3.1 1. Folded Morgan Fingerprints + CACTVS Fingerprints

The Morgan fingerprint algorithm, developed by Morgan (1965) and later refined by Rogers and Hahn, generates circular substructure features around each atom in a molecule.
For an atom a at iteration i, the identifier is computed as:

$$ID_i(a) = \text{hash}\left(\text{sort}\left(\{ID_{i-1}(n) : n \in \mathcal{N}(a)\}\right)\right)$$

where $\mathcal{N}(a)$ represents the set of neighboring atoms of atom a. A hash function h maps arbitrary input data to fixed-size integers: $h : \{0,1\}^* \to \{0,1\}^n$, where molecular substructures are converted to reproducible numerical identifiers for fingerprint generation.

The folding process maps variable-length feature sets to fixed-length binary vectors:

$$bit\_position = \text{hash}(feature) \bmod N$$

where N is the desired fingerprint length (1024 in our case and is also the standard practice).

CACTVS (Computer Aided Chemistry Training and Visualization System) fingerprints complement Morgan fingerprints by encoding additional structural features like functional group patterns, ring systems and aromaticity, pharmacophore features, and topological descriptors.
The combined fingerprint is expressed as:

$$\mathbf{F}_{combined} = \mathbf{F}_{Morgan(folded)} \oplus \mathbf{F}_{CACTVS}$$

where $\oplus$ denotes concatenation operation.

Bit Collisions in Folded Fingerprints
Bit collisions occur when multiple distinct molecular features hash to the same bit position in a folded fingerprint. This phenomenon can lead to false positives in similarity calculations.
For a fingerprint of length N bits and k features, the probability of at least one collision is:

$$P(\text{collision}) = 1 - \frac{N!}{(N-k)! \cdot N^k}$$

For large N and moderate k, this approximates to:

$$P(\text{collision}) \approx 1 - e^{-\frac{k(k-1)}{2N}}$$

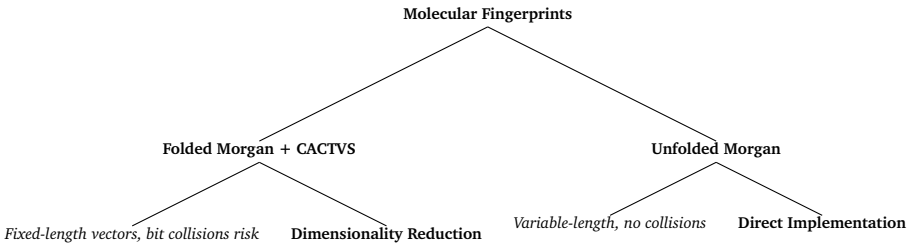
The equations are derived using the fact that the bit collision concept is analogous to the birthday paradox[23], and the $Poisson$ distribution is used for large N.

8.3.2 2. Unfolded Morgan Fingerprints

Unfolded Morgan fingerprints preserve the original hash values without mapping them to a fixed-length bit vector. This approach eliminates bit collisions, but results in variable-length representations.
An unfolded fingerprint is represented as:

$$\mathbf{F}_{unfolded} = \{h_1, h_2, \ldots, h_k\}$$

where each h_i is a unique hash value corresponding to a circular substructure.
For our project, we proceeded to use both the types of fingerprints in a manner detailed in the schema given below:



8.3.3 Practical Example of bit collision

Consider a molecule with the following circular substructures:

- Feature A: `C-C-C` (hash: 12345)
- Feature B: `N-C-O` (hash: 67890)

For a 1024-bit fingerprint:

$$\text{Position A} = 12345 \bmod 1024 = 345 \tag{1}$$

$$\text{Position B} = 67890 \bmod 1024 = 345 \tag{2}$$

Both features map to bit 345, creating a collision.
The false positive rate due to collisions can be estimated as:

$$FPR = \frac{\text{Number of collided bits}}{\text{Total number of set bits}}$$

8.3.4 Comparative analysis of fingerprints

Aspect	Folded Morgan	Unfolded Morgan	Scaffolded
Memory Usage	Fixed (low)	Variable (high)	Fixed (medium)
Collision Risk	High	None	Medium
Similarity Speed	Fast	Medium	Fast
Information Loss	Yes	No	Partial
Interpretability	Low	High	High
Scalability	Excellent	Poor	Good

Table 14: *Comparison of Fingerprint Approaches*

8.3.5 Scaffolded Fingerprints

A molecular scaffold represents the core structural framework of a molecule, typically obtained by removing side chains and reducing complex ring systems to their fundamental topology.
The Murcko scaffold extraction algorithm identifies the largest ring system and connecting chains:

$$Scaffold(M) = \text{RemoveSideChains}(\text{ExtractRingSystem}(M))$$

8.4 TruncatedSVD vs. SparsePCA

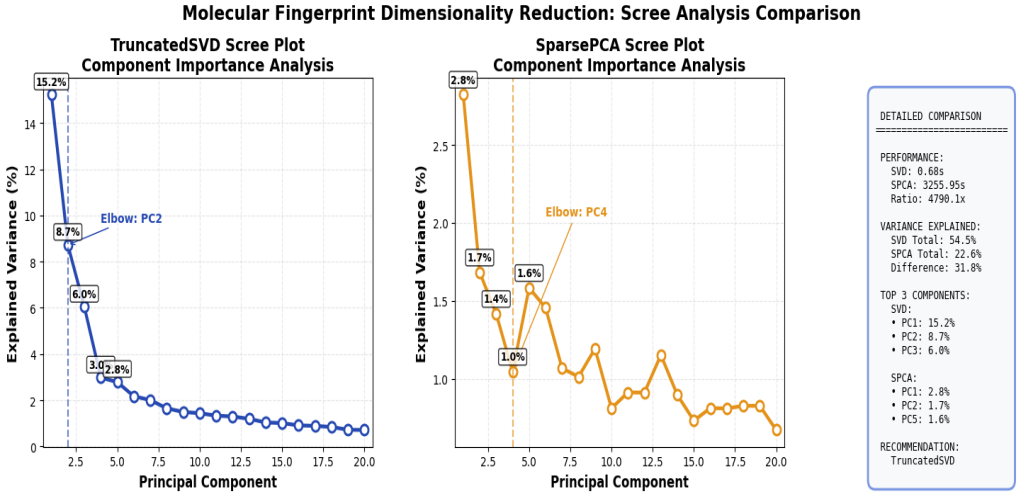


Figure 24: *Comparison between SparsePCA and TruncatedSVD*

It is noteworthy that in addition to a much better explained variance with fewer components, the amount of time to fit TruncatedSVD was **0.68 seconds** and for SparsePCA was **3255.95 seconds**, making TruncatedSVD **more than 4700 times faster**.

8.5 Clustering using K-Means and DBSCAN

For the visualisation of the clusters, we tried using K-Means. We had tested k (number of clusters) values from 2-10, selecting k=3 based on the highest silhouette score (0.6014). However, since our clusters had ambiguous shapes, the results were not up to the mark, as possible outlier points were forced to be a part of a cluster, and we could not move forward with even a single cluster and create a cluster-specific model.

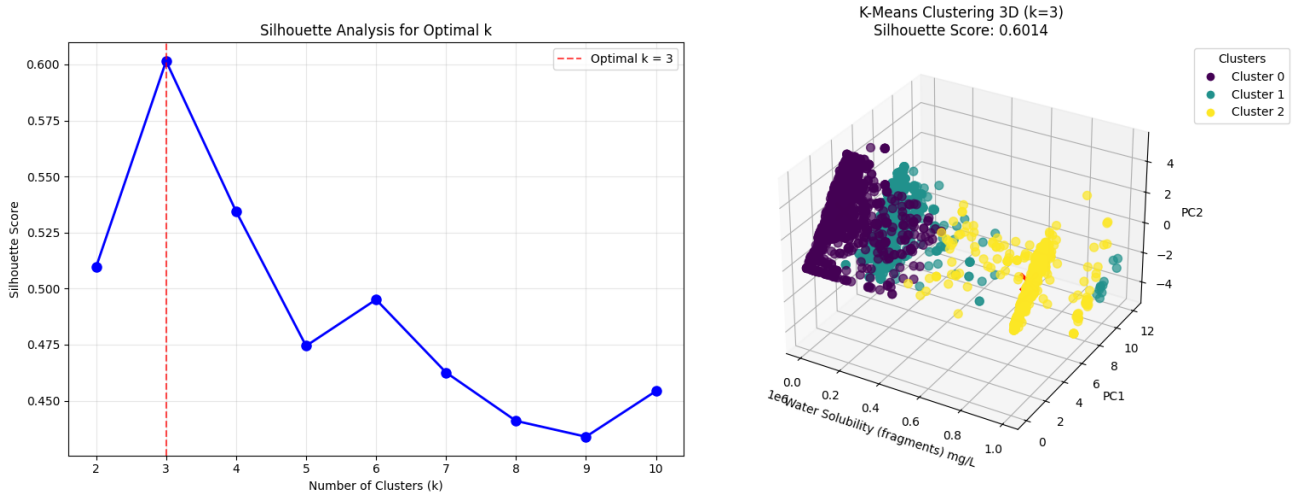


Figure 25: *Clustering with KMeans*

This was followed by use of DBSCAN. The reason being that it does well with ambiguously shaped clusters. However, we had to optimize two parameters, epsilon (eps) which describes the maximum

distance between two points for them to be considered neighbors in the same cluster, and minimum number of samples (`min_samples`) which is the minimum number of points (including the point itself) required in a neighborhood to form a dense core point that can start a cluster. For the first estimate of `min_samples`, we used the rule of thumb by using twice the number of fingerprint columns (after dimension reduction). We used k-distance graph to get the first estimate of `eps` (using the elbow), then grid-searched `eps` and `min_samples` combinations using silhouette scoring. The search space implemented for the grid search for both parameters were expanded by taking values near the estimated ones.

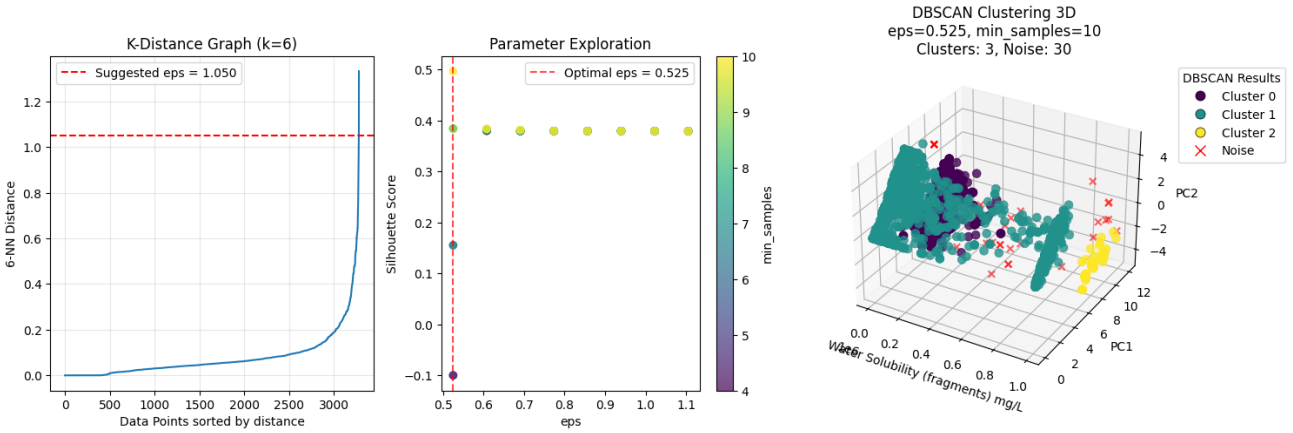


Figure 26: *Clustering with DBSCAN*

It does not perform well with clusters with varying densities, as it assumes uniform density across all clusters and is highly sensitive to parameter selection. This is visible from the results observed, as cluster 1 could be broken into two separate clusters based on observation. In the K-distance graph which is plotted using `NearestNeighbours()`, we can see that the suggested `eps` calculated using the `np.argmax(differences)` is higher than the optimal value found during the optimization, making it a necessary task during cluster formation analysis.