From Structures to Risk: p53-Mediated Carcinogenicity Prediction of Atmosphere Chemicals with Machine Learning

Department of Chemistry

MSc Data Science Project report

From Structures to Risk: p53-Mediated Carcinogenicity Prediction of Atmosphere Chemical with Machine Learning

prepared by

JASHWANTHIKAA DURAIKANNU RAJENDIRAN

Project Supervisor Prof. Jacqui Hamilton ……

Date 28/08/2025 ……

No. of Words 6112 ……

2024/25 Entry Cohort

Department of Chemistry: MSc Projects

Declaration of Help Received

Name of Student……JASHWANTHIKAA DURAIKANNU RAJENDIRAN……

Project Supervisor(s)……Prof. JACQUI HAMILTON……

This form should be completed by each student and submitted as part of the report. The information provided allows proper acknowledgement of the help you have received and will help the markers in evaluating your performance.

1. Did anyone other than the project supervisor(s) assist you in carrying out your project? Tick the appropriate box:

   ✓ ☐ Yes ☐ No

   If yes, detail the help received (Suggested categories are given below the table):

Name of Helper	Category (i to v)	Details (if necessary)
Thomas Cornell	(i)	

(i) Day to day help
(ii) Initial training in a technique or operation
(iii) Specific assistance (e.g. operation of instrumentation)
(iv) Help with interpretation of results
(v) Any other category of help

2. Did your project supervisor(s) provide you with material assistance towards completing your project (e.g. code which produced or analysed results, graphs which you have included in your report)? Tick the appropriate box:

   ✓ ☐ Yes ☐ No

   If yes, please provide details of the help received:

My supervisor helped me provide code for SMILES Fingerprint Calculation, assisted with SIRIUS prediction setup, and supported access to the real laboratory sample used for validation.

3. I have read the Assessment Information in the MSc Handbook on academic misconduct and plagiarism. I have completed the Academic Integrity Tutorial.
   I declare that this assessment is a presentation of original work, I am the sole author and I am not attempting to pass off work created by generative AI content tools as my own.

Name of student (in lieu of signature)

Jashwanthikaa Duraikannu Rajendiran
……28/08/2025……

Abstract

Natural and Human-made aerosol are atmospheric particles that have the potential to carry carcinogenic chemicals that can harm public health. Conventional tests for toxicity tests, although well established, are time-consuming, expensive, and limited by ethics (can't test on people and must limit animal use), and thus unsuitable for large-scale environmental screening. This study solves two primary goals: building a model for predicting p53-mediated carcinogenicity of ambient chemicals and validating it against independent spectral measurements and real laboratory samples. To fulfil these aims, the study applies the XGBoost machine learning model that has been trained on dataset-repaired assay data of the U.S. EPA CompTox Chemicals Dashboard (ATG_p53_CIS) with cheminformatics-derived molecular fingerprints and monoisotopic mass as predictive variables.

Data were extensively pre-processed for chemical quality and relevance before model training. Validation was initially performed with liquid chromatography–mass spectrometry (LC-MS) spectra from MassBank, which were processed in SIRIUS software to produce structure-based predictions for direct comparison of model output. The model was then transferred to LC-MS data of laboratory-collected aerosol-related samples for real-world usability testing.

The results show that the combination of high-quality chemical datasets, interpretable machine learning, and staged validation provides an efficient, scalable, and ethical solution to carcinogenicity evaluation. The solution improves aerosol toxicity prediction, enhances the integration of computational tools in environmental health, and enables future regulatory and monitoring systems to ensure public health protection.

Table of Contents

Project report

1. Introduction

1.1 Latent Cancer Toxicity and Air Pollution

The most detrimental environment health risk to human beings is air pollution, estimated to have caused around seven million premature fatalities every year (WHO, 2023). Most troubling is fine particulate matter with a diameter smaller than 2.5 microns ($PM_{2.5}$) since it can enter the body's own natural air defences, go deep into tissue within the lungs and into the blood (Brook et al., 2010; Pope III and and Dockery, 2006). If in the blood, they can cause one to be more susceptible to disease (Kim et al., 2015).

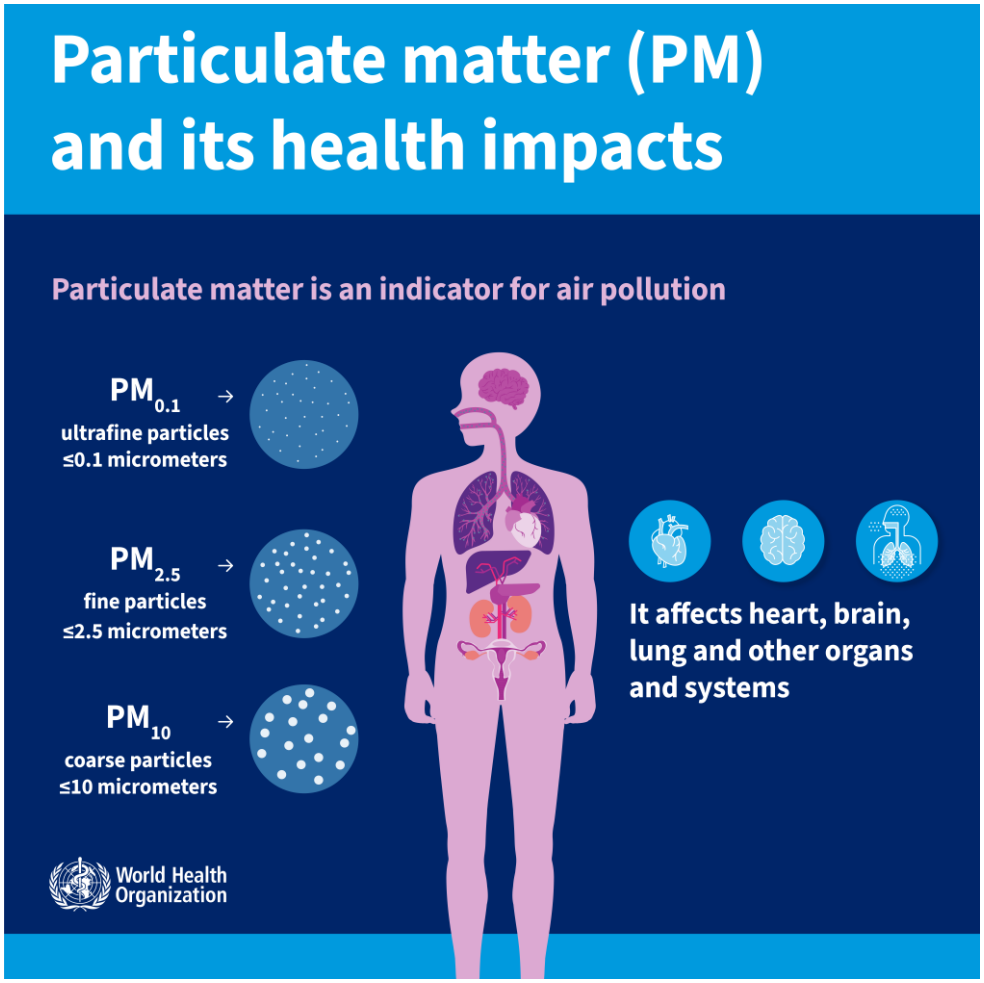

Figure 1.1. Schematic of $PM_{2.5}$ particle penetration into the respiratory tract and bloodstream, highlighting pathways of systemic health effects (WHO, 2025).

The danger posed by $PM_{2.5}$ is not just because of its very small size but also because of the complex chemical composition of these particles. They have a natural tendency to transfer a mix of the toxicants including polycyclic aromatic hydrocarbons (PAHs), volatile organic compounds (VOCs), nitroaromatics and transition metals such as lead and cadmium, together with reactive oxygen species (ROS), all of which have well-documented connections to

carcinogenic processes (Kelly and Fussell, 2012; Kovacic and Somanathan, 2014; Shrivastava et al., 2017). They are primarily derived from automotive exhausts, fossil fuel combustion, industrial discharge, and agricultural burning of biomass (Hallquist et al., 2009; Pöschl, 2005; Seinfeld and Pandis, 2016).

Due to its deep lung penetration and chemical toxics load, $PM_{2.5}$ is also widely reported to cause a series of chronic diseases like cardiovascular and respiratory diseases, and importantly, a series of cancers (Eaves et al., 2020; Khan et al., 2022; Li et al., 2018; Lima de Albuquerque et al., 2021; Thangavel et al., 2022). In the face of these comparatively more well documented risks, fewer than five percent of airborne chemicals are thoroughly tested, long-term toxicological assessment (Cao et al., 2025; Li et al., 2011; Thomas et al., 2019). Inadequate toxicological data hinder the ability of public health agencies and regulatory agencies to respond appropriately to emerging environmental risks.

1.2 p53 Pathway and Cancer

Cancer results from step-wise accumulation of mutations inactivating the body's defence mechanisms, permitting unlimited growth, immune evasion, and genomic injury-induced survival (Hanahan and Weinberg, 2011). Central to such defences is the p53 protein, whose gene product is TP53, a tumour suppressor transcription factor (Lane, 1992; Levine, 1997).

In normal conditions, p53 remains low. When there is DNA damage or cell stress, however, p53 is rapidly activated and can halt the cell cycle to facilitate repair, initiate DNA repair processes, or induce apoptosis (programmed cell death) to eliminate irreparably damaged cells. These functions, as shown in Figure 1.2 (Kastenhuber and Lowe, 2017), depict the "guardian of the genome" role of p53 in guiding cells back to stability or safe removal.

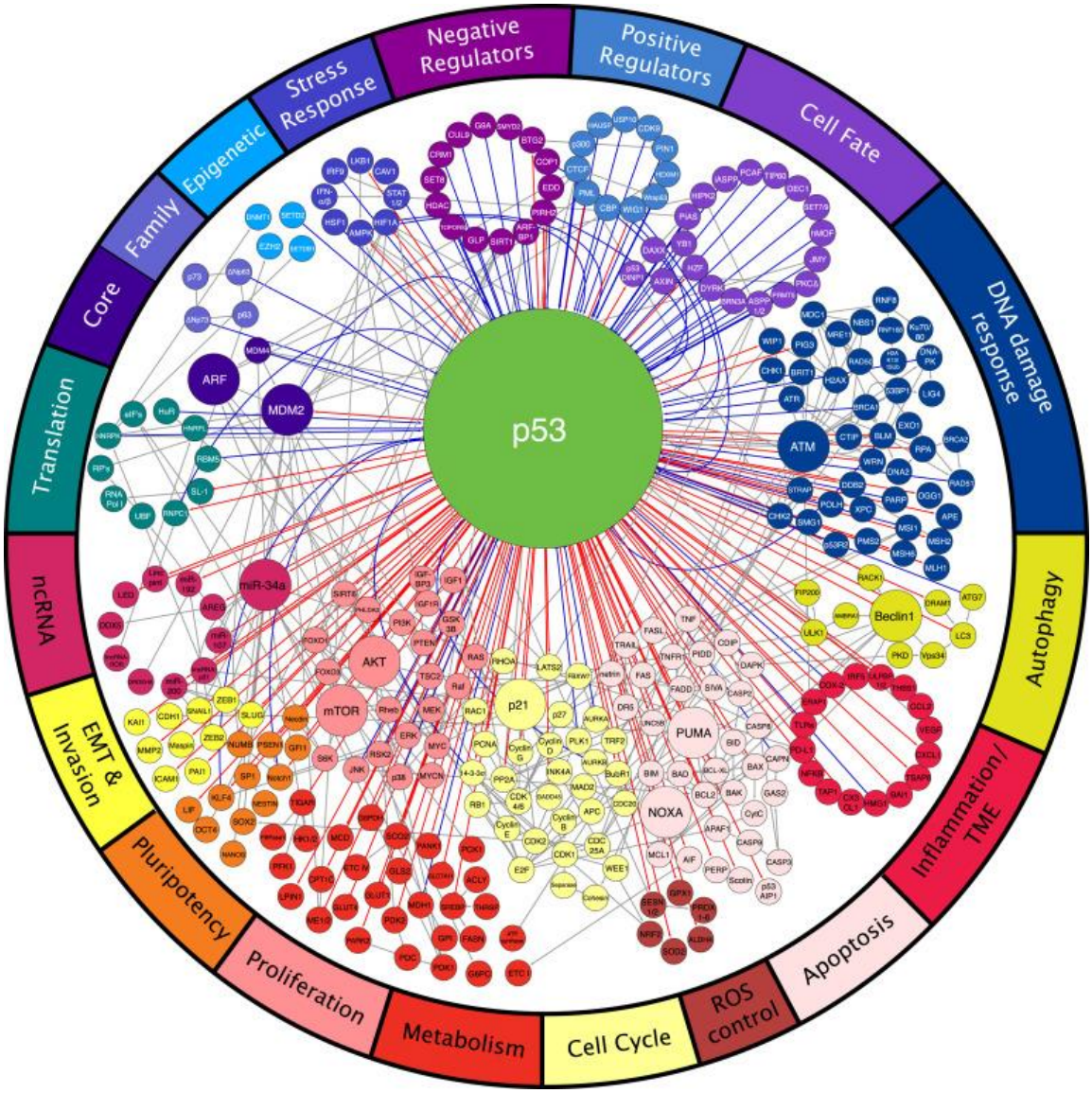

Figure 1.2. The p53 network.
p53 acts like a master switch: it receives many signals (blue lines) and, in response, turns many genes on or off (red lines), which then talk to each other (gray lines). Arrows represent "activates," T-bars represent "blocks," and solid dots represent "modifies." (Kastenhuber and Lowe, 2017).

Because of its central position, TP53 mutations represent among the most common human cancer alterations in more than half of all cases (Kastenhuber and Lowe, 2017; Olivier et al., 2010). Significantly, environmental carcinogens such as harmful chemicals in dirty air have the ability to disrupt this pathway, either through inducing DNA damage or influencing p53 function (Perwez Hussain and Harris, 2007; Zeron-Medina et al., 2013). This makes p53 not only a central tumour suppressor but a mechanistic biomarker for the identification of carcinogenic exposures, for example, from air pollutants.

1.3 Traditional and Computational Approaches to Chemical Toxicity Evaluation

Animal experiments have been the gold standard method of carcinogenicity evaluation with rat and mouse chronic bioassays as the classical legacy to note tumour growth following chemical exposure (Krewski et al., 2010; Zeiger, 2019). While these tests have provided valuable reference data, they are time- and resource-demanding, costly, and pose ethical concerns due to high animal use. Most importantly, animal models generally will not be capable of simulating human-restricted metabolic pathways and genetic vulnerabilities, hence limiting their applicability for human health assessment (Basketter et al., 2012).

Classic testing also tests chemicals in isolation, while humans are exposed to complex mixtures. Consequently, vast regions of chemical space remain untested and underspecified. The U.S. Environmental Protection Agency (EPA) CompTox Chemicals Dashboard contains over 875,000 compounds out of which a minor fraction have solid toxicological data (Williams et al., 2017). To fill this knowledge gap, New Approach Methodologies (NAMs) have been developed that include high-throughput in vitro tests, organ-on-a-chip devices, and mechanistic assays based on well-defined Adverse Outcome Pathway (AOP) key events. NAMs offer faster, scalable, and more responsible assessments together with mechanistic insight (Escher et al., 2022; OECD, 2016).

Whereas this, computational toxicology moved away from Quantitative Structure–Activity Relationships (QSAR) and Quantitative Structure–Property Relationships (QSPR) towards machine learning approaches. QSAR/QSPR models are solid but bounded by domains of applicability, complexity bifurcation at new compounds or blends in the lack of analogues, and they break down with external validation issues (OECD, 2023). Previous machine learning models such as logistic regression, support vector machines (SVMs), and random forests increased the strength of prediction while struggling against high-dimensional, heterogeneous data. Sophisticated deep models offer better performance at the cost of "black box" interpretability, raising the demand for interpretability to grant regulatory clearance (Jia et al., 2023; Mayr et al., 2016).

Widening convergence of NAMs and computational methods can be observed in modern toxicology: omic and high-throughput data are being used more and more in predictive models, basically taking NAM ability beyond experimental throughput limits (Najjar et al., 2023; Thomas et al., 2019).

1.4 Data Science Meets Toxicology: Predicting Chemical Risk with Machine Learning

Machine learning (ML) is increasingly the focus of computational toxicology as it permits predictions of chemical toxicity based on molecular structure, physicochemical descriptors, and high-throughput assay data. This in silico toxicology offers scalability and speed so that large libraries of chemicals can be screened without the cost or ethics of animal testing (Cavasotto and Scardino, 2022; Wu et al., 2022).

Of machine learning methods, the gradient boosting algorithm XGBoost has been especially successful. It is efficient computationally, handles missing data, and handles class imbalance, which is common in toxicology data (Chen and Guestrin, 2016; Kang et al., 2023). Combined

with interpretability tools such as SHAP (Shapley Additive Explanations), XGBoost models not only make precise predictions but also specify which molecular descriptors define toxicity outcomes (Lundberg and Lee, 2017; Ponce-Bobadilla et al., 2024). This is important because clear models help construct regulatory trust: agencies will be more willing to act upon ML predictions if rationales that support them are understandable and scientifically motivated (Jia et al., 2023).

Recent studies illustrate the growing use of ML in toxicology. (Peets et al., 2022) developed MS2Tox, a predictor of ecotoxicity based on MS^2 fragmentations tandem mass spectrometry data that capture structural breakdown patterns of molecules. (Dührkop et al., 2019) demonstrated that SIRIUS and CSI:FingerID can predict molecular structures and activity from MS/MS spectra, extending toxicity prediction even to chemicals whose identity is unknown. (Arturi and Hollender, 2023) used Random Forests in risk-guided prioritization of environmental contaminants, and (Jaganathan et al., 2022) constructed an interpretable respiratory toxicity model from optimal molecular descriptors.

Outside of toxicity, ML has also been applied. (Palm and Kruve, 2022) showed it applied to unknowns' quantitation in LC/HRMS workflows, while (Born et al., 2023) integrated ML into digital discovery platforms for chemical investigation. Graph Neural Networks (GNNs) are advanced techniques that can be trained from molecular graphs directly, which is a direction of great hope for future toxicology (Rong et al., 2020).

Together, these studies demonstrate how ML, from ensemble models to deep learning, and most recently to GNNs, is transforming toxicology with predictive, interpretable, and scalable assessments that benefit regulatory screening as well as hypothesis generation.

1.5 Bridging Prediction and Reality: Validation Using Mass Spectrometry Data

Predictive models are meaningful only if they reflect real-world chemical exposures, not curated datasets. Environmental toxicology has generally employed liquid chromatography–tandem mass spectrometry (LC-MS/MS), distinguishing between molecules and disassembling them into ions to provide very high-resolution chemical fingerprints. These methods are crucial to Secondary Organic Aerosol (SOA) analysis, which is of great importance to generating fine particulate matter ($PM_{2.5}$) and is a well-established health problem (Shrivastava et al., 2017; Witkowski and Gierczak, 2017). The majority of SOA-derived chemicals remain to be characterised, with incomplete coverage in regulatory databases, and LC-MS/MS is instrumental in expanding the observable chemical world.

Fragmentation patterns in tandem mass spectrometry (MS^2 or MS/MS) hold structural information that can be interpreted through computation. Software such as SIRIUS and CSI:FingerID are classic examples of this procedure in constructing molecular formulas and predicting substructural fingerprints from spectra (Dührkop et al., 2019, 2015). Subsequent results may then be integrated into machine learning models, with the spectral fingerprints being used to predict potential toxicological attributes. For example, MS2Tox demonstrates

that ecotoxicity may be predicted with XGBoost on MS²-derived fingerprints to extend risk assessment even to previously unknown compounds (Peets et al., 2022).

Building on such methods, this study will explore whether LCSB MassBank LC-MS/MS data (Elapavalore et al., 2023) could be interpreted using spectral interpretation software to generate fingerprints, which in turn could be utilized to inform predictive carcinogenicity models. By doing this, the study aims to align lab-based predictions better with environmental exposures as they happen, while also continuing to facilitate the application of computational toxicology in chemical risk assessment.

1.6 Why This Research Matters: Safer Air Through Smarter Models

The air pollutant mix contains thousands of chemicals, but very few have been tested to see if they are carcinogens (Thomas et al., 2019). Traditional animal testing is time-consuming, expensive, and ethically questionable (Basketter et al., 2012; Zeiger, 2019). Newer methods, such as high-throughput screening (HTS) and computer toxicology, can screen more rapidly, in volume, and with minimal reference to animal studies (Jaworska and Hoffmann, 2010; OECD, 2016)

The aim of this work is to build a machine learning model that can predict atmospheric chemicals' carcinogenicity based on molecular fingerprints and p53 assay data. The model provides a faster, scalable alternative to traditional toxicology methods. A machine learning algorithm, XGBoost, is employed in this research to predict whether a chemical will or will not interfere with a cancer-related biological pathway, from its molecular "fingerprints" and high-throughput screening information. Molecular fingerprints are quantitative descriptors that represent the structure of a molecule in a computer-processable way (Guha, 2007). High-throughput assays are laboratory tests that are computer controlled and quickly assay the number of chemicals that have an effect on a specific biological process (Kavlock et al., 2008).

To validate the model for real-world use, the model is also compared to analytical measurements of environmental samples conducted through tandem mass spectrometry (MS/MS). Computational models like CSI:FingerID predict the chemical structures based on spectral patterns, similar to air monitoring research of pollutants.

This strategy underpins significant safety and ethics models, such as Next Generation Risk Assessment (NGRA) and the 3Rs principles of Replacement (wherever possible avoiding the use of animals), Reduction (minimising the number of animals used) and Refinement (reducing the harm caused by tests) (Russell and Burch, n.d.; Thomas et al., 2019). It is made to be reproducible and flexible in order for regulators and scientists to more accurately determine high-risk air pollutants with higher efficiency, particularly in industrially or traffic-influenced regions (Arturi and Hollender, 2023; Landrigan et al., 2018).

2. Data, Tools and Machine Learning Methodology

This subsection provides a detailed explanation of the data sources, cheminformatics feature engineering, machine learning strategy, and external validation protocols employed to

construct a robust and interpretable predictive model of chemical carcinogenicity. The broad goal of this research is to predict the likelihood of an atmospheric organic compound activating the p53 tumour suppressor pathway, indicative of genotoxic and carcinogenic activity. The pipeline integrates publicly available high-throughput toxicological screens with cheminformatics-based structural descriptors, advanced machine learning algorithms, and external spectral fingerprint verification. Analyses were conducted using R version 4.3.1 on an open-source, reproducible computing environment.

2.1 Biological Reasoning and Data Collection

The primary source of data for the study was the EPA CompTox Chemicals Dashboard ("CompTox Chemicals Dashboard," n.d.), which aggregates high-throughput screening (HTS) data from the ToxCast and Tox21 initiatives. Specifically, the ATG_P53_CIS assay was used that identifies transcriptional activity of tumour suppressor protein p53. The protein is implicated in DNA repair, cell cycle arrest, and apoptosis and is also referred to as the "guardian of the genome" (Levine, 1997).
Mutations or disruptions in p53 are linked to over 50% of all human cancers (Vousden and Lane, 2007) and hence represent a significant biomarker for toxicological testing.

From the CompTox dataset, an initial list of 4039 compounds was drawn with active and inactive labels that describe whether they are able to activate a signal in the ATG_P53_CIS assay. Only organic compounds with valid SMILES strings, CAS numbers, and monoisotopic mass were retained. Further filtering was carried out for the elimination of duplicates, mixtures, and compounds with undefined molecular formulae. To make it relevant to atmospheric chemistry, a further filter was used to keep only those compounds with C, H, N, O, and S atoms, which are typical constituents of airborne organics (Shrivastava et al., 2017).

After missing-data entry removal and extensive structural and assay-based filtering, the final cleaned dataset of 718 compounds, consisting of 127 actives and 591 inactives, was obtained. They were the basis of all the subsequent cheminformatics processing and model construction.

2.2 Cheminformatics Feature Generation

In order to convert chemical structures to computational inputs, a wide range of molecular fingerprints were computed from the rcdk package in R that encapsulates the Chemistry Development Kit (CDK) (Guha, 2007; Steinbeck et al., 2003). Fingerprints are numerical or binary descriptors that characterize substructural motifs, topology, and physicochemical features of molecules.

A range of fingerprints was employed for providing high-density structural data. They included MACCS keys (166-bit substructure motifs) (Durant et al., 2002), PubChem fingerprints (881-bit descriptors used in PubChem database) (Wang et al., 2009), and Klekota–Roth fingerprints (4860-bit motifs tuned for bioactivity prediction) (Klekota and Roth, 2008). Other descriptors reflecting electronic and topological properties, such as CDK standard fingerprints (1024-bit) (Steinbeck et al., 2003) and EState fingerprints (Hall and Kier, 1995), were also generated.

All fingerprint vectors were normalized, near-zero variance features removed, and descriptors with high correlation (> 0.9) removed. All this preprocessing is crucial to stability and generalizability in cheminformatics models (Svetnik et al., 2003).

2.3 Data Preprocessing and Feature Selection

Preprocessing was conducted in R with the caret (Kuhn, 2008) and dplyr (Wickham, n.d.) packages that have peer-reviewed data cleaning, manipulation, and model preparation structures. Missing values were removed from observations so that complete cases could be used. Continuous descriptors such as monoisotopic mass were z-score normalized so that all features are on the same scale and so as not to enable large-valued variables to dominate model learning.

To enhance model stability, predictors of near-zero variance were removed by the use of caret's nearZeroVar function because such predictors contribute almost nothing to classification. Furthermore, high pair-wise correlated features (Pearson's $|r| > 0.9$) were also removed to reduce redundancy and fight collinearity-cause instability. While such operations discard some of the predictive data, they improve interpretability and yield a more stable subset of single predictors (Guyon and Elisseeff, n.d.; Yvan, 2007). The benefits of this filtering are evident from the correlation heatmap of the top 20 features (Fig. 2.3.1), where numbers of highly correlated variables were reduced to a minimum so that independent descriptors could make more balanced contributions.

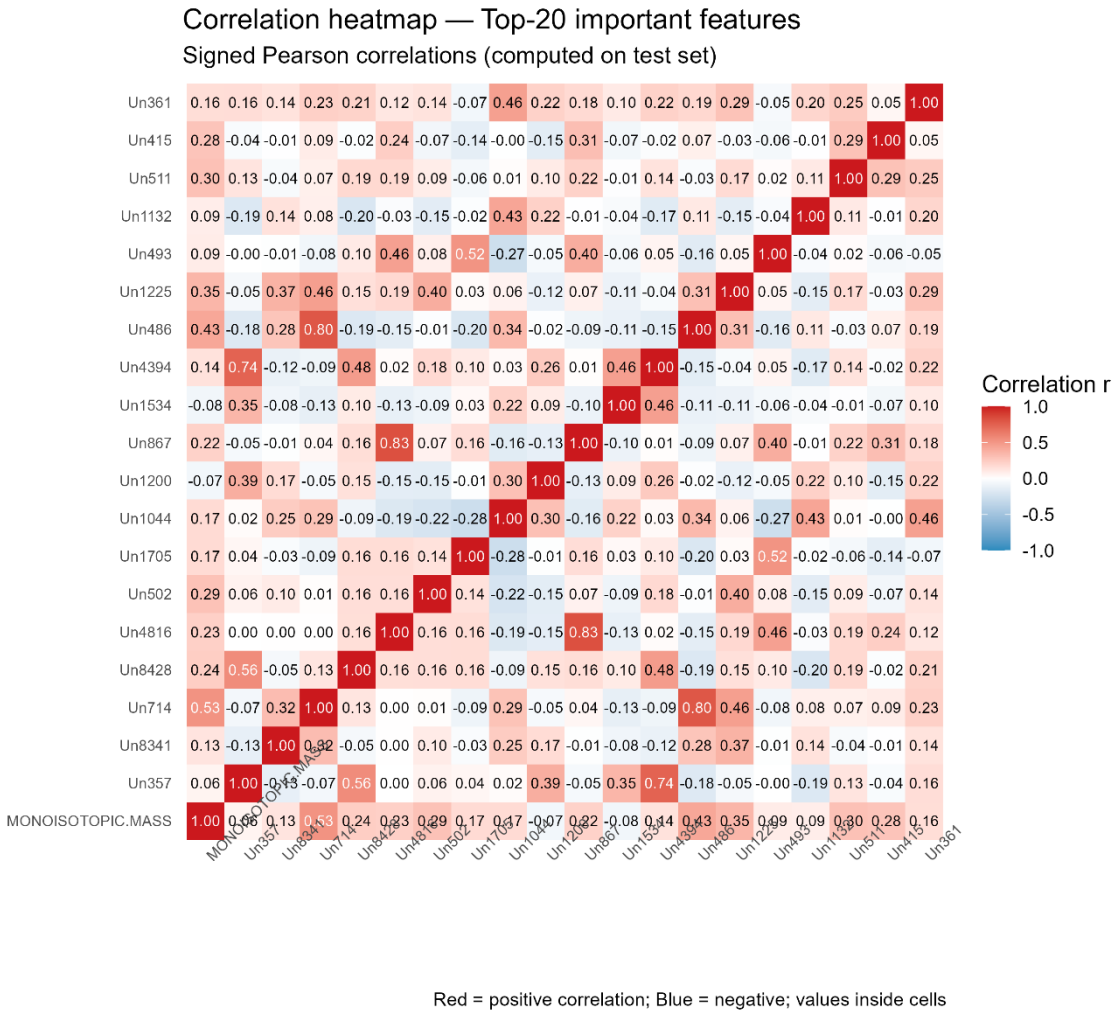

Figure 2.3.1. Correlation heatmap of the top-20 most important features, showing Pearson correlations computed on the test set. Red cells indicate positive correlation, blue cells negative correlation, with values shown inside the grid.

The binary outcome variable HIT.CALL was also numerically coded (active = 1, inactive = 0). Class imbalance was managed by applying stratified train/test split (80:20 split) and XGBoost scaling of the scale_pos_weight parameter, which reformulated the loss function as a function of class distribution. Surprisingly, weighting filled in for but did not replace stratification for proportional representation of minority actives. Finally, the pre-processed data was stored in the xgb.DMatrix sparse matrix data format to ensure maximum memory usage and computational efficiency during training large models.

2.4 XGBoost Machine Learning Model Development

For predicting carcinogenic activity based on structural fingerprints, the present study employed the Extreme Gradient Boosting (XGBoost) algorithm, a decision-tree-based ensemble model popular for its high reliability and efficiency in high-dimensional space (Chen and Guestrin, 2016). Rather than shallow benchmarking of various algorithms, the work attempted to extensively optimize and validate XGBoost, facilitating conservative calibration, interpretability, and external validation.

XGBoost constructs boosted decision tree ensembles iteratively reducing errors from previous iterations. The algorithm is defined by the use of second-order gradient approximation to optimize a smooth and differentiable loss function for more convergence with no loss in accuracy (Chen and Guestrin, 2016; Ke et al., 2017). L1 (Lasso) and L2 (Ridge) regularization penalties prevent overfitting, while row and column subsampling prevents overdependence on certain features or samples. A decreasing learning rate (eta) sacrifices speed of convergence and generalisation, and the gamma parameter prevents too rapid a loss reduction before a split, in order to maintain tree complexity justified. These characteristics are especially suitable to cheminformatics data, where descriptors are of high dimensionality, correlated, and potentially redundant (Probst et al., 2019; Walter, 2022).

Building the model first utilized 10-fold cross-validation within the caret R package but was computationally intensive while accuracy gains were slight. As a result, the protocol was minimized to 5-fold cross-validation, a setting demonstrated satisfactory for toxicology modelling and significantly reducing runtime (Boulesteix et al., 2017). Instead of tuning all the hyperparameters simultaneously, models were trained across single boosting rounds (nrounds) to monitor classification dynamics. Performance stabilized at ~100 rounds, with further iterations making diminishing returns (Fig. 2.4.1).

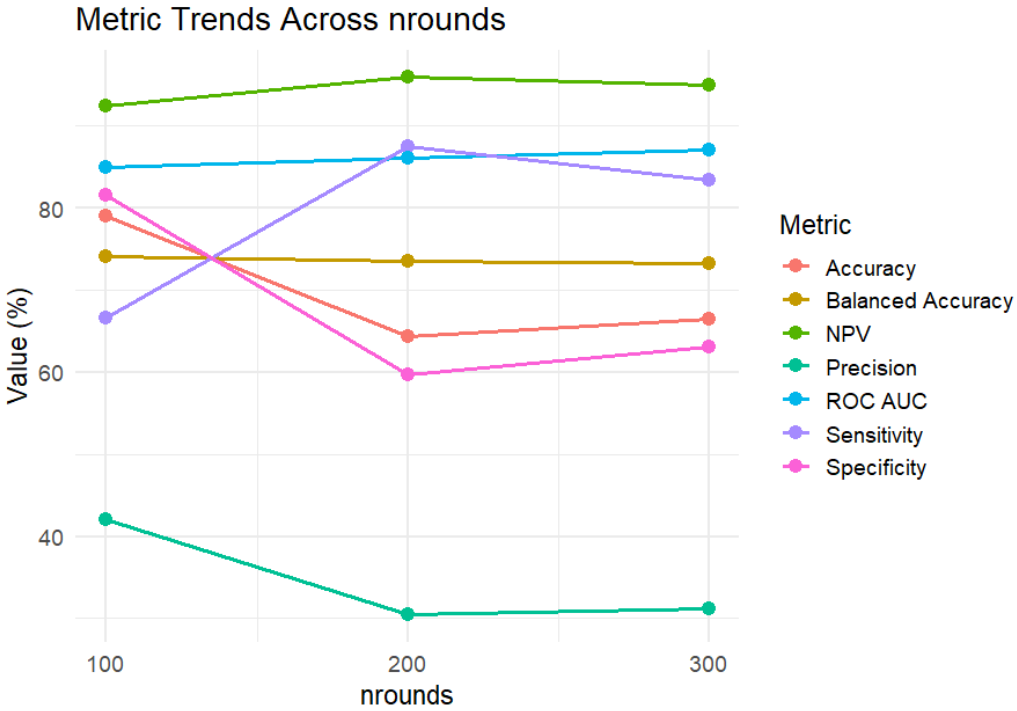

Figure 2.4.1. Trends in classification metrics across different boosting rounds (nrounds).

The optimal parameters in the training set were nrounds = 100, eta = 0.01, max_depth = 6, gamma = 5, colsample_bytree = 0.6, subsample = 0.6. This model was then retrained on the quality-controlled 718-compound data set (127 actives, 591 inactives) after initial filtering and quality control had ensured chemical relevance and descriptor stability (Section 2.3). Re-training on the full curated set of data allowed the algorithm to utilize all the information

available, a method recommended for small to moderate-sized toxicological datasets (Kuhn and Johnson, 2013).

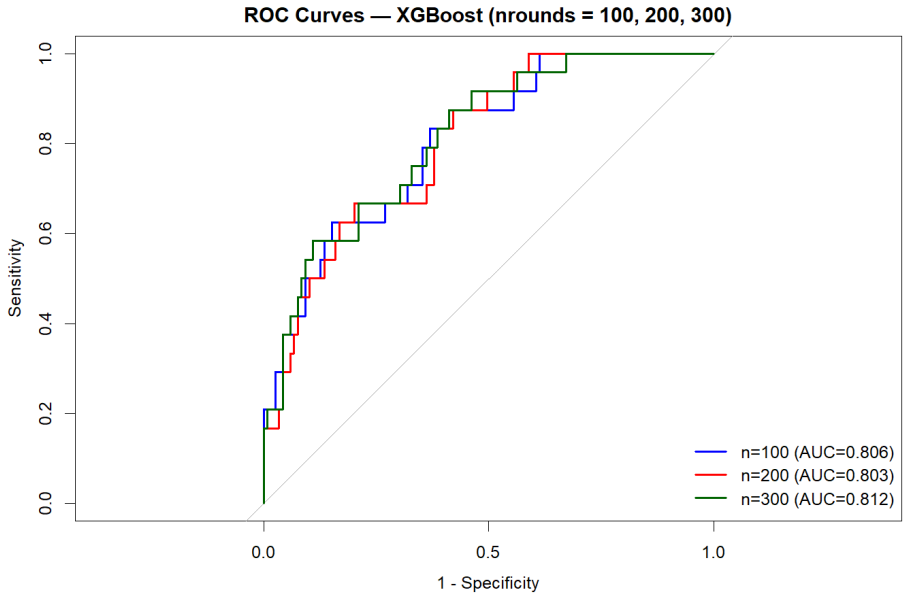

Figure 2.4.2. Comparative bar plot of performance metrics for models trained with nrounds = 100, 200, and 300.

Threshold optimisation was guided by Receiver Operating Characteristic (ROC) analysis, with Youden's J statistic used to balance sensitivity and specificity (Powers, 2020; Youden, 1950). Figure. 2.4.2, which were trained at nrounds = 100, 200, and 300, showed minimal divergence after over 100 rounds, confirming that this was a viable alternative. As threshold = 0.47, sensitivity rose to 92.91% (118/127 actives detected), but false positives rose to 164, reducing precision to 41.84%. This configuration is preferred whenever the cost of failing to detect a carcinogen exceeds false alarms, such as in population screening for public health (Chawla et al., 2002). With a threshold of 0.65, false positives were reduced to 40 and raised specificity (93.23%) and precision (65.22%) but reduced sensitivity to 59.06%. supplies. This trade-off, on the threshold optimisation curve (Fig. 2.4.3), places the higher threshold more appropriately in regulatory or poverty settings where false positives are extremely expensive (Hastie et al., 2009).

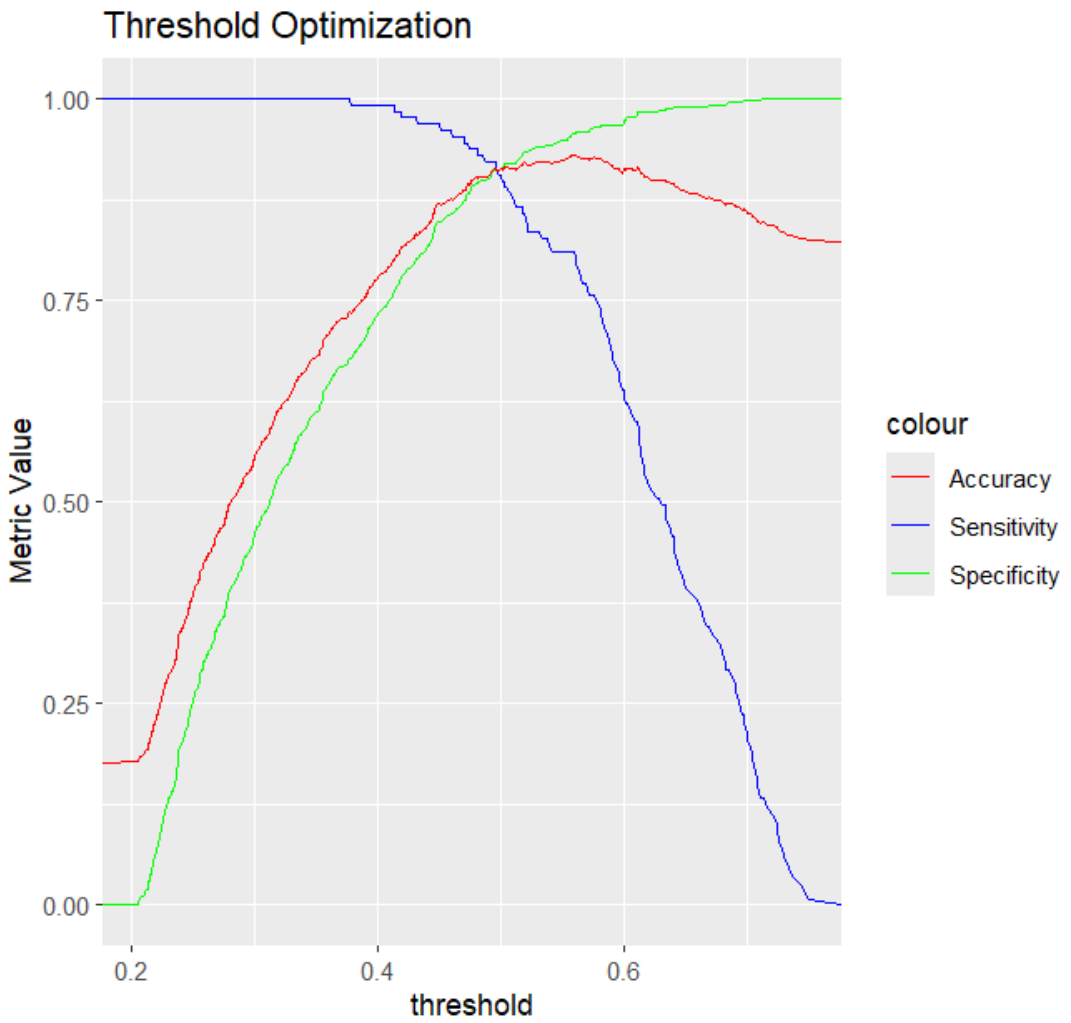

Figure 2.4.3. Threshold optimization curve showing trade-offs between sensitivity, specificity, and accuracy for the final model.

The final retrained model, tuned at a ROC-derived cutoff of 0.4805, was 90.25% accurate, 93.70% sensitive, 89.51% specific, and 91.61% balanced accurate. The Kappa statistic (0.7131) further attested to an agreement more than by chance (Brodersen et al., 2010; Cohen, 1960). The ROC curve (Fig. 2.4.4) also reflected an AUC value of greater than 0.90, denoting the strong discrimination power of the model between active and inactive compounds.

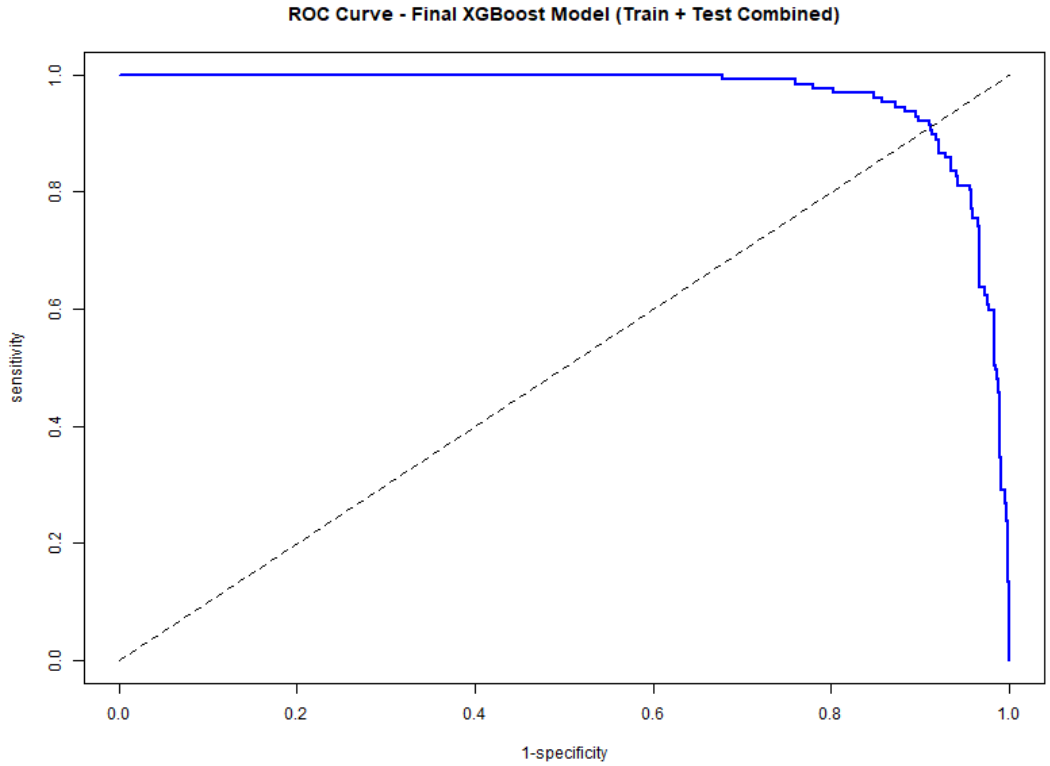

Figure 2.4.4. Receiver Operating Characteristic (ROC) curve of the final XGBoost model retrained on the full dataset with optimised hyperparameters. The model demonstrates strong discriminatory ability between active and inactive compounds, with an AUC indicating robust generalisation across both classes.

All the performance measures using ROC analysis, confusion matrices, and calibration were conducted with pROC and caret packages of R. Reproducibility was ensured by fixing the random seeds and saving the trained objects using saveRDS, following best practice in computational modelling and open science (Sandve et al., 2013; Wilkinson et al., 2016).

Cumulatively, this workflow demonstrates the methodological strength and toxicological appropriateness of an optimised XGBoost model, with suitably aimed optimisation. By sensitivity versus precision balancing by thresholding, the model is tuned across different use cases from precautionary screening to regulatory prioritisation. This conjunction of machine learning and toxicological thinking is the overall movement towards computational alternatives facilitating traditional assays and read-across strategies (Sheridan et al., 2016).

2.5 External Validation Based on Mass Spectral Data

External validation plays an important role in determining whether a machine learning model will be able to generalise from controlled training data and make good predictions in real-world scenarios. Without testing, models would be overfitted to optimized descriptors, making them less applicable in practice and regulation. To achieve this, the XGBoost classifier which was trained was validated using high-resolution tandem mass spectrometry (MS/MS) data for the Luxembourg Centre for Systems Biomedicine (LCSB) MassBank. The repository is applied

extensively in pipelines of non-target screening, particularly where chemical structure is unknown or partial (Kind and Fiehn, 2010; Schymanski et al., 2014).

There were over 2,200 compounds in the raw dataset. There was a lot of curations required in order to select for quality: those compounds that had no identifiers, sets of metadata that were incomplete, or no experimentally derived structures were excluded. Duplicates were merged on the elemental formula level, as typically performed in non-target analysis (Schymanski et al., 2014). Cross-referencing with toxicological databases then left 319 compounds with labels for p53 activity, making up the final validation set (Judson et al., 2010; Richard et al., 2016). This reduction indicates the extent of harmonisation and quality control that is needed to get good external datasets.

Spectral fingerprints were generated from MS/MS fragmentations by CSI:FingerID in SIRIUS v5.6.3 (Dührkop et al., 2019). As opposed to cheminformatics descriptors calculated directly from molecular structures, these fingerprints are calculated in a direct fashion from spectra, which introduces what is referred to as a domain shift a systematic gap between the training and test feature space distributions (Quinonero-Candela et al., 2022; Sugiyama et al., n.d.). To assist in counterbalancing this, spectral features were mapped to the model development descriptor schema and the same preprocessing steps (z-score scaling, near-zero variance filtering, correlation filtering) applied to the features to maintain consistency.

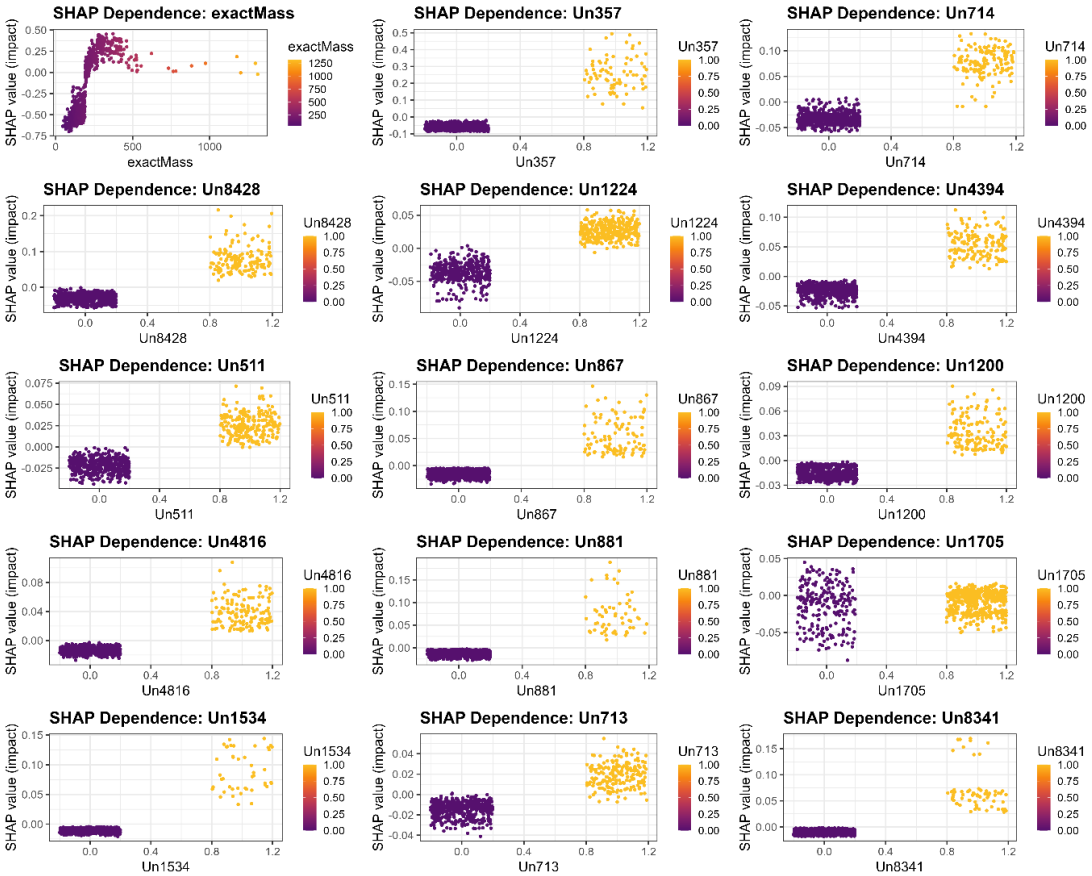

Figure 2.5.1. SHAP dependence plots for the top 15 features (e.g., exact mass, Un714, Un8428), showing how increasing or decreasing feature values (colour-coded by intensity) influence carcinogenicity predictions.

Feature interpretability was verified using SHAP dependence plots for the 15 most informative variables (Figure 2.5.1). These revealed that higher monoisotopic mass values were more likely to make positive contributions to carcinogenicity prediction, and binary fingerprints such as Un714 and Un8428 display clear presence absence effects with strong impacts on results. This confirmed that even during domain shift, the model retained chemically interpretable decision logic.

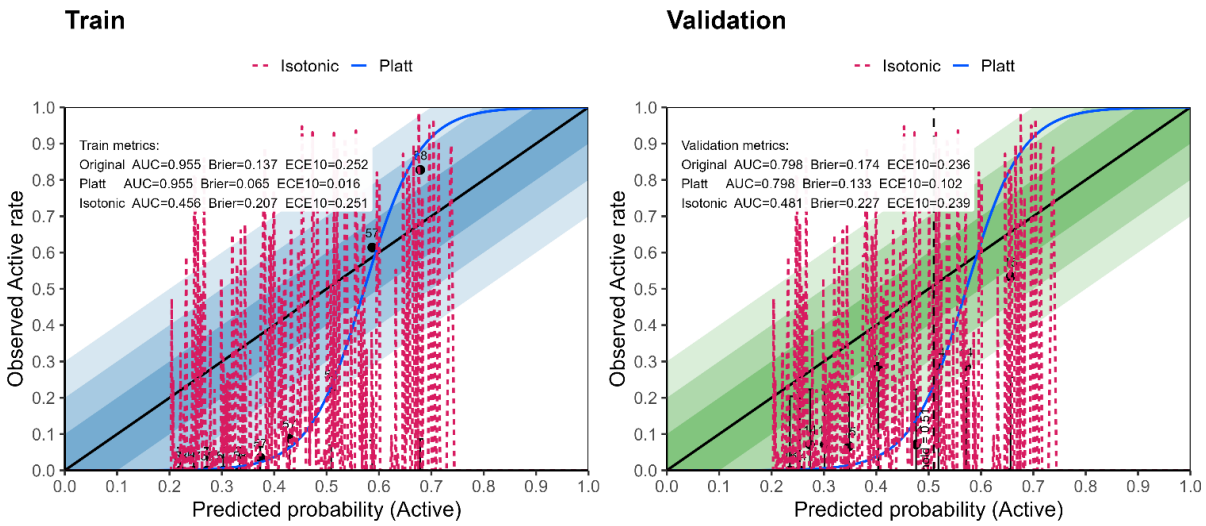

Figure 2.5.2. Calibration plots showing that Platt scaling improved probability reliability over isotonic regression, with lower Brier scores and ECE.

Tests of calibration also measured reliability of probability. Platt scaling and isotonic regression are contrasted with the uncalibrated model in Figure 2.5.2. Platt scaling, based on logistic mapping from raw scores, reduced the Brier score and expected calibration error (ECE) to deliver better-calibrated probabilities (Guo et al., 2017; Niculescu-Mizil and Caruana, 2005). Isotonic regression, by contrast, delivered evidence of overfitting the relatively small validation set, corroborating Platt scaling as the superior correction.

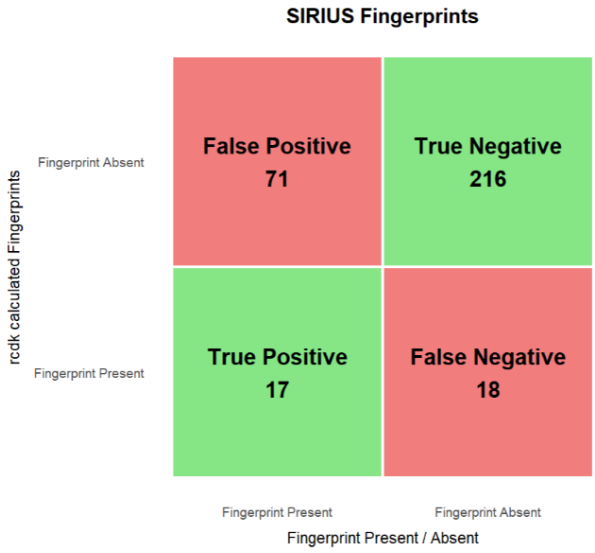

Figure 2.5.3. Confusion matrix and contingency table of predicted vs. actual activity classes in the external validation dataset.

The performance results are listed in the confusion matrix (Figure 2.5.3). The model was 72.36% correct: 17 carcinogens were correctly predicted (true positives) and 18 were left out (false negatives), and 216 inactives were correctly predicted (true negatives) and 71 were incorrectly predicted to be carcinogens (false positives). These results are equivalent to 48.6% sensitivity, 75.3% specificity, and balanced accuracy of 62.0%. Bootstrapped resampling (1,000 replicates) placed the general accuracy between 67–77%, which indicates robustness in the presence of the small size of the dataset. In contrast to the ~91% balanced accuracy in Section 2.4 for carefully curated fingerprints, the degradation highlights the cost of domain shift. This level of reduction is, nonetheless, within the earlier studies seen within the literature, whereby cross-validation over spectral fingerprints would typically log 70–75% accuracies (Dührkop et al., 2019; Peets et al., 2022; Walter, 2022).

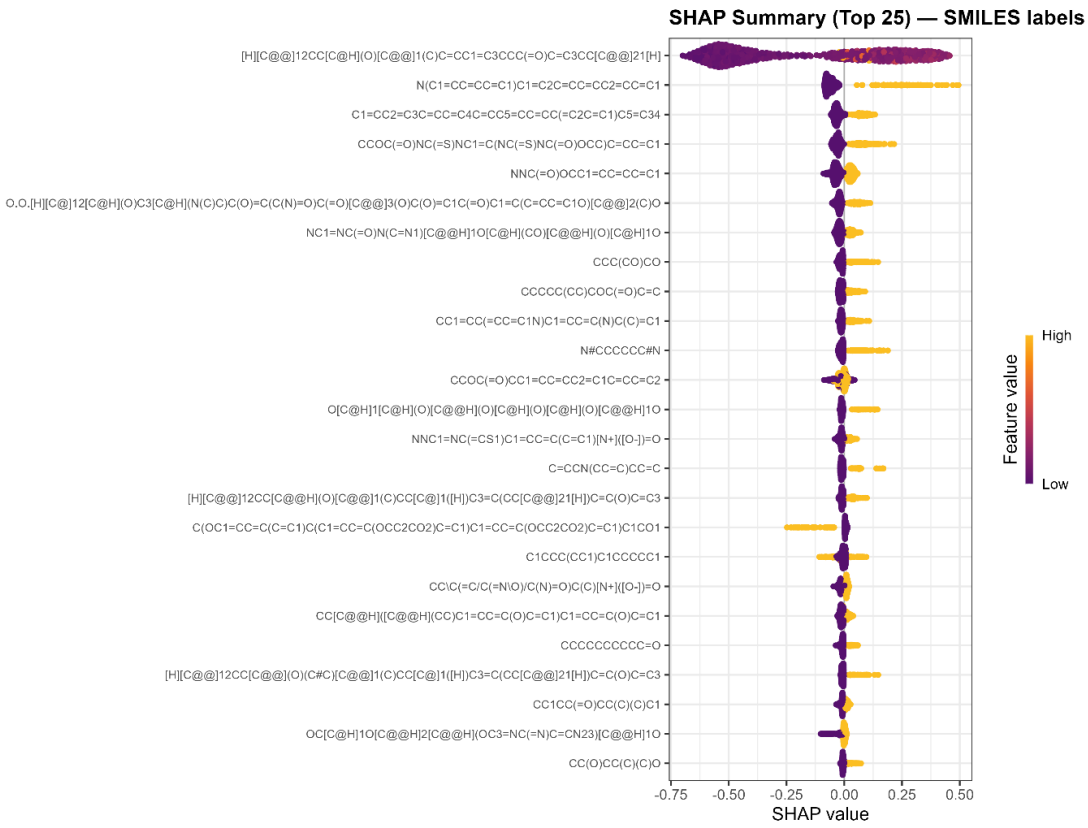

Figure 2.5.4 SHAP beeswarm summary plot with SMILES annotations for the top 25 features, linking structural motifs to predicted carcinogenicity.

In order to place predictions into context, SHAP beeswarm plots with SMILES annotation were generated for the most significant 25 features (Figure 2.5.4). These showed substructures such as aromatic rings, nitro groups, and dense heteroatom moieties being most frequently associated with predictions of carcinogenicity, in accordance with known toxicophores. This illustrates how sparse fingerprints can be translated into chemical features that are understandable.

Generally, this was externally validated with external validation and demonstrated that the XGBoost classifier was able to maintain its predictability under domain shift upon training on

spectrum-derived fingerprints with ~72% accuracy and 62% balanced accuracy. While accuracy was lower than with curated descriptors, this was consistent with literature benchmarks (Dührkop et al., 2019; Helma et al., 2004; Walter, 2022). These findings provide evidence for the prospects of predictive toxicology to be applied to analytically derived but structurally uncertain chemicals and to complement high-throughput risk prioritisation separate from animal studies.

3. Results: Model Deployment to Actual Laboratory SOA Samples

In this section of the study, it was determined if and how the machine learning classifier might be deployed from controlled samples in the lab to actual atmospheric chemical mixtures, in this case, secondary organic aerosols (SOA). Most SOA studies proceed only as far as chemical characterization, but in this case the intent was to see if a human-health toxicology model could have some knowledge of atmospheric oxidation products' carcinogenicity. This is one of the first attempts to bridge atmospheric chemistry with predictive toxicology.

3.1 Sample Generation and Analysis

Limonene ($C_{10}H_{16}$) was chosen as the precursor, an organic volatile compound emitted by plants and also present in everyday consumer items like air fresheners and spray cleaners. Its chemical composition has two double bonds (Figure 3.1), which are accountable for highly reacting with ozone. These positions are accountable for the elevated SOA yields when limonene gets oxidized in the atmosphere (Bateman et al., 2009; Hallquist et al., 2009).

The experiments were conducted in the NCAS/University of York atmospheric chamber (Pereira et al., 2019). Limonene was mixed with ozone under controlled conditions to mimic real atmospheric oxidation. Teflon-filtered particles were solvent-extracted and analysed by LC–MS/MS. The process in the chamber ensured that the measured oxidation products were actual atmospheric processes and not theoretical chemistry only (Witkowski and Gierczak, 2017).

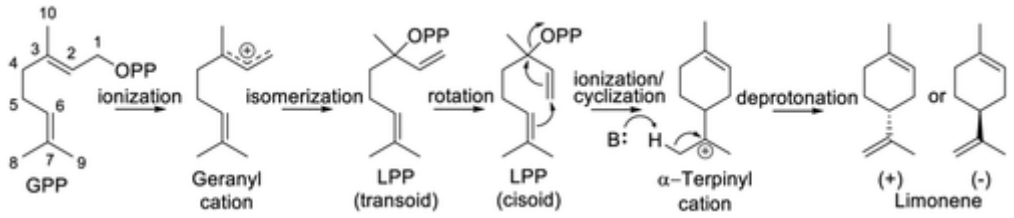

Figure 3.1. Limonene structure ($C_{10}H_{16}$) with its two reactive double bonds being responsible for its rapid oxidation and strong SOA formation (Morehouse et al., 2017).

3.2 Formula Attribution and Fingerprint Inference

LC–MS/MS spectra were processed using SIRIUS v5.6, which determines molecular formulas from exact mass and isotope pattern, and substructural fingerprints were predicted through CSI:FingerID (Böcker and Dührkop, 2016; Dührkop et al., 2019). 506 oxidation products were noted out of which 500 were annotated at high confidence level with formulas.

The compounds had monomers (single-molecule units, typically C_9–C_{10}), dimers (two joined units, for instance, C_{18}–C_{20}), and higher size oligomers (lots of joined units). Because of isomerism (different arrangements of atoms that share the same formula), some formulas such as $C_{10}H_{16}O_4$ appeared multiple times at different retention times, which were structural isomers.

This was performed to transform raw spectral data to machine learning model-needed fingerprint features.

3.3 Model Predictions

The XGBoost classifier was then trained using the 500 labelled oxidation products. The probability of each compound to be Active in activating the p53 cancer pathway was predicted and converted into Active or Inactive calls utilizing the ROC-optimal threshold (0.4805). The number of compounds predicted Active and Inactive respectively from the 500 compounds were 107 (21.4%) Active and 393 (78.6%) Inactive. That is, roughly one in five compounds derived from limonene exhibited carcinogenic activity, a proportion comparable to that discovered in high-quality toxicology collections like ToxCast (Judson et al., 2010).

There were evident structural patterns. Fewer than 10% of C_{10} or smaller monomers were predicted Active, while about a quarter of the dimers and over 30% of the oligomers were designated as Active. This is evidence that larger oxygen-contained molecules have a greater predicted biological reactivity, which corresponds with toxicological theory that multifunctional, oxidised organics will bind to DNA and proteins (Schwöbel et al., 2011).

3.4 Literature Validation

For consistency, the following annotated formulas were validated against earlier chamber studies (Witkowski and Gierczak, 2017). Frequently occurring small molecules such as $C_{10}H_{16}O_4$ and $C_9H_{14}O_4$ were correctly predicted Inactive, while less frequent, oxygen-rich compounds such as $C_{18}H_{28}O_6$ and $C_{19}H_{28}O_7$ were determined to be predicted Active. Figure 3.2 illustrates this dichotomy, with on the left the frequency with which reported formulas are seen in the literature and on the right the predicted carcinogenicity outcome. The lesson is that frequency of detection does not equate to toxicological importance: occasional oligomers may pose higher predicted carcinogenicity risk than frequent small molecules.

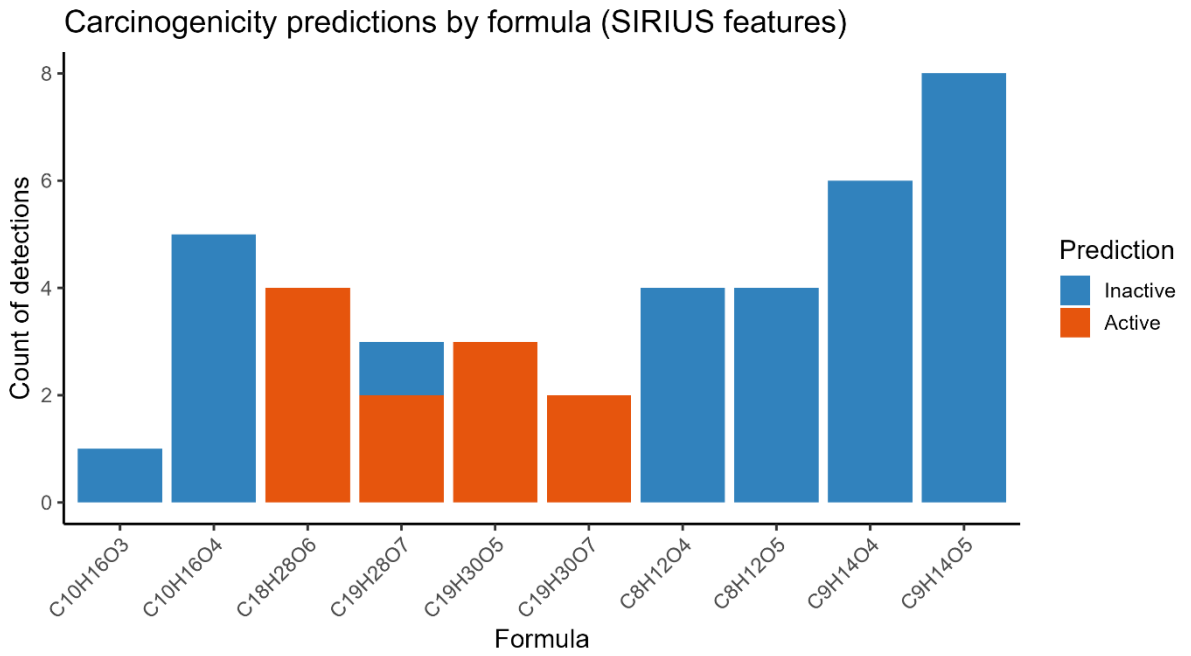

Figure 3.2. Detection frequency of literature-reported limonene SOA formulas in SIRIUS-processed LC–MS/MS data alongside their predicted carcinogenicity (Active or Inactive) based on SIRIUS-derived molecular fingerprints using the trained XGBoost classifier.

3.5 Significance and Limitations

Unlike Section 2.5, which focused on external validation with curated spectral libraries, this section demonstrates deployment on real, uncurated mixtures where uncertainty and domain shift are much greater. While absolute validation is not possible without bioassays, the model successfully prioritised oxygenated oligomers as high-risk candidates, narrowing down hundreds of detected products into a manageable subset for experimental follow-up (Lazic and Williams, 2021; Sobus et al., 2018).

These findings present a proof-of-concept for applying computational toxicology to atmospheric blends. Through chamber exposures integration, LC–MS/MS quantitation, and predictive modelling, this technique enables the identification of potentially carcinogenic aerosol sources. This technique can be applied to other monoterpenes, such as α-pinene and β-pinene, which would broaden the application to indoor as well as outdoor air quality.

4. Discussion

We examined if machine learning could yield a mechanistically informed, reliable framework to predict the carcinogenic potential of ambient organic chemicals. Results indicate an XGBoost classifier trained using p53 assay data had high performance on curated datasets and generalized well to spectral fingerprints and secondary organic aerosol mixtures. This is among the first demonstrations of the direct application of predictive toxicology to atmospheric chemistry, connecting chemical characterisation with health-relevant risk assessment.

Performance on curated CompToX data highlighted the utility of cheminformatics features. Structural fingerprints combined with monoisotopic mass recorded p53 activation patterns with

balanced accuracies >90% at ROC-optimised thresholds. Extended to MS/MS-derived fingerprints, accuracy fell to ~72%. This reduction is indicative of the lower information content of spectra compared to intact structures but remains comparable to other external validation studies (Dührkop et al., 2019; Peets et al., 2022). Rather than indicating model weakness, this performance demonstrates robustness to domain shift, where training and test data differ in feature representation.

Deployment to limonene SOA provided distinct structural trends. Fewer than 10% of small monomers (C_{10} or fewer carbons) were predicted Active, compared to ~25% of dimers and more than 30% of oligomers. Larger oxygen-rich multifunctional organics were therefore more likely to be predicted Active, in agreement with mechanistic expectation and previous toxicological evidence (Schwöbel et al., 2011). That such trends emerged from a model trained solely on toxicology data testifies to its explanatory power for atmospheric mixtures.

Several caveats must be mentioned. The CompTox dataset is imbalanced with relatively few active compounds, which even with weighting schemes can produce biased results. Only XGBoost was examined in depth; while extremely well-suited to high-dimensional data, comparison with Random Forests, Support Vector Machines, or graph neural networks would provide more scope for generalisability (Limbu and Dakshanamurthy, 2022; Mayr et al., 2016). External validation was restricted to MassBank data, which covers only a fraction of atmospheric chemical space. Also, fingerprint inference from spectra is ambiguous, particularly for mass-resolution-indistinguishable isomers.

Despite these constraints, the model is demonstrated to be of practical use as a prioritisation tool. With less than 5% of air chemicals currently profiled toxicologically (Thomas et al., 2019), the capacity to rank compounds in silico is valuable. The approach reduces the need for animal testing, is aligned with New Approach Methodologies (OECD, 2016), and provides a scalable solution for the identification of high-risk candidates for targeted laboratory testing.

Future studies need to expand both chemical and methodological scope. Adding more mechanistic assays than p53 would follow multiple carcinogenic pathways. Expanding the workflow to other SOA precursors, such as α-pinene and β-pinene, and to atmospheric ambient air samples would enhance robustness. Adding tools for interpretability like SHAP would further connect predictions back to structural motifs, increasing linkage to mechanistic toxicology (Lundberg and Lee, 2017; Walter, 2022).

As a conclusion, this study illustrates that machine learning can advance predictive toxicology from curated datasets to real-world atmospheric mixtures. The innovation is in illustrating the direct applicability of a human-health toxicology model to SOA. More broadly, the workflow provides a scalable, ethical, and mechanistically informed platform for chemical risk assessment in air quality research.

5. Conclusion

In this project, a machine learning model was developed to forecast atmospheric chemicals' carcinogenicity with the mechanistic endpoint of p53 pathway activation. By training an

XGBoost model on CompTox assay data, the study showed that structural fingerprints and monoisotopic mass features are competent to identify toxicity-related patterns. The model proved to be strongly internally accurate and stable on testing against spectrum-derived fingerprints, claiming its ability to generalise beyond curated sets. Application to limonene secondary organic aerosol also indicated that the framework is able to provide environmentally relevant information, with more oxygenated and higher molecular weight compounds being consistently predicted as more likely carcinogens.
These results demonstrate how computational toxicology can progress toward real-world applicability by linking in vitro assays, cheminformatics descriptors, and air mixture. The framework provides a scalable and ethical approach to ranking hazardous chemicals, aligned with modern practices of chemical risk assessment minimizing animal usage.

However, this paper is also sensitive to its own limitations. Imbalance between the dataset of active and inactive compounds, reliance on a single algorithm, and absence of overt biological validation for SOA predictions undermine the confidence that can be placed in the conclusions. These are problems that look forward to future directions that include using multiple assays, varied machine learning models, and interpretation methods to learn more about model decision-making.

In total, this project demonstrates that machine learning can contribute meaningfully to the study of atmospheric carcinogens, both in a scientific and in a potential public health protection sense.

Data Access

Project code and processed data are openly available in the GitHub repository:
https://github.com/jashwanthikaa/Predicting-Carcinogenicity.git

In addition, all project code and datasets have also been archived in the supplementary file Predicting Carcinogenicity.zip, which has been submitted through the VLE. This ensures long-term accessibility even if the online repository is updated.

Bibliography

Arturi, K., Hollender, J., 2023. Machine Learning-Based Hazard-Driven Prioritization of Features in Nontarget Screening of Environmental High-Resolution Mass Spectrometry Data. Environ. Sci. Technol. 57, 18067–18079. https://doi.org/10.1021/acs.est.3c00304

Basketter, D.A., Clewell, H., Kimber, I., Rossi, A., Blaauboer, B., Burrier, R., Daneshian, M., Eskes, C., Goldberg, A., Hasiwa, N., Hoffmann, S., Jaworska, J., Knudsen, T.B., Landsiedel, R., Leist, M., Locke, P., Maxwell, G., McKim, J., McVey, E.A., Ouédraogo, G., Patlewicz, G., Pelkonen, O., Roggen, E., Rovida, C., Ruhdel, I., Schwarz, M., Schepky, A., Schoeters, G., Skinner, N., Trentz, K., Turner, M., Vanparys, P., Yager, J., Zurlo, J., Hartung, T., 2012. A roadmap for the development of alternative (non-animal) methods for systemic toxicity testing. ALTEX - Altern. Anim. Exp. 29, 3–91. https://doi.org/10.14573/altex.2012.1.003

Bateman, A.P., Nizkorodov, S.A., Laskin, J., Laskin, A., 2009. Time-resolved molecular characterization of limonene/ozone aerosol using high-resolution electrospray ionization mass spectrometry. Phys. Chem. Chem. Phys. 11, 7931. https://doi.org/10.1039/b905288g

Böcker, S., Dührkop, K., 2016. Fragmentation trees reloaded. J. Cheminformatics 8, 5. https://doi.org/10.1186/s13321-016-0116-8

Born, J., Markert, G., Janakarajan, N., B. Kimber, T., Volkamer, A., Rodríguez Martínez, M., Manica, M., 2023. v. Digit. Discov. 2, 674–691. https://doi.org/10.1039/D2DD00099G

Boulesteix, A.-L., Wilson, R., Hapfelmeier, A., 2017. Towards evidence-based computational statistics: lessons from clinical research on the role and design of real-data benchmark studies. BMC Med. Res. Methodol. 17, 138. https://doi.org/10.1186/s12874-017-0417-2

Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010. The Balanced Accuracy and Its Posterior Distribution, in: 2010 20th International Conference on Pattern Recognition. Presented at the 2010 20th International Conference on Pattern Recognition, pp. 3121–3124. https://doi.org/10.1109/ICPR.2010.764

Brook, R.D., Rajagopalan, S., Pope, C.A., Brook, J.R., Bhatnagar, A., Diez-Roux, A.V., Holguin, F., Hong, Y., Luepker, R.V., Mittleman, M.A., Peters, A., Siscovick, D., Smith, S.C., Whitsel, L., Kaufman, J.D., 2010. Particulate Matter Air Pollution and Cardiovascular Disease. Circulation 121, 2331–2378. https://doi.org/10.1161/CIR.0b013e3181dbece1

Cao, F., Zhao, X., Fu, X., Jin, Y., 2025. Computational insights into exploring the potential effects of environmental contaminants on human health. Sci. Rep. 15, 11779. https://doi.org/10.1038/s41598-025-96193-2

Cavasotto, C.N., Scardino, V., 2022. Machine Learning Toxicity Prediction: Latest Advances by Toxicity End Point. ACS Omega 7, 47536–47546. https://doi.org/10.1021/acsomega.2c05693

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. 16, 321–357. https://doi.org/10.1613/jair.953

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Presented at the KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco California USA, pp. 785–794. https://doi.org/10.1145/2939672.2939785

Cohen, J., 1960. A Coefficient of Agreement for Nominal Scales. Educ. Psychol. Meas. 20, 37–46. https://doi.org/10.1177/001316446002000104

CompTox Chemicals Dashboard [WWW Document], n.d. URL https://comptox.epa.gov/dashboard/ (accessed 8.20.25).

Dührkop, K., Fleischauer, M., Ludwig, M., Aksenov, A.A., Melnik, A.V., Meusel, M., Dorrestein, P.C., Rousu, J., Böcker, S., 2019. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat. Methods 16, 299–302. https://doi.org/10.1038/s41592-019-0344-8

Dührkop, K., Shen, H., Meusel, M., Rousu, J., Böcker, S., 2015. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proc. Natl. Acad. Sci. 112, 12580–12585. https://doi.org/10.1073/pnas.1509788112

Durant, J.L., Leland, B.A., Henry, D.R., Nourse, J.G., 2002. Reoptimization of MDL Keys for Use in Drug Discovery. J. Chem. Inf. Comput. Sci. 42, 1273–1280. https://doi.org/10.1021/ci010132r

Eaves, L.A., Smeester, L., Hartwell, H.J., Lin, Y.-H., Arashiro, M., Zhang, Z., Gold, A., Surratt, J.D., Fry, R.C., 2020. Isoprene-Derived Secondary Organic Aerosol Induces the Expression of MicroRNAs Associated with Inflammatory/Oxidative Stress Response in Lung Cells. Chem. Res. Toxicol. 33, 381–387. https://doi.org/10.1021/acs.chemrestox.9b00322

Elapavalore, A., Kondić, T., Singh, R.R., Shoemaker, B.A., Thiessen, P.A., Zhang, J., Bolton, E.E., Schymanski, E.L., 2023. Adding open spectral data to MassBank and PubChem using open source tools to support non-targeted exposomics of mixtures. Environ. Sci. Process. Impacts 25, 1788–1801. https://doi.org/10.1039/D3EM00181D

Escher, S.E., Partosch, F., Konzok, S., Jennings, P., Luijten, M., Kienhuis, A., de Leeuw, V., Reuss, R., Lindemann, K.-M., Bennekou, S.H., 2022. Development of a Roadmap for Action on New Approach Methodologies in Risk Assessment. EFSA Support. Publ. 19, 7341E. https://doi.org/10.2903/sp.efsa.2022.EN-7341

Guha, R., 2007. Chemical Informatics Functionality in R.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On Calibration of Modern Neural Networks.

Guyon, I., Elisseeff, A., n.d. An Introduction to Variable and Feature Selection.

Hall, L.H., Kier, L.B., 1995. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. J. Chem. Inf. Comput. Sci. 35, 1039–1045. https://doi.org/10.1021/ci00028a014

Hallquist, M., Wenger, J.C., Baltensperger, U., Rudich, Y., Simpson, D., Claeys, M., Dommen, J., Donahue, N.M., George, C., Goldstein, A.H., Hamilton, J.F., Herrmann, H., Hoffmann, T., Iinuma, Y., Jang, M., Jenkin, M.E., Jimenez, J.L., Kiendler-Scharr, A., Maenhaut, W., McFiggans, G., Mentel, T.F., Monod, A., Prévôt, A.S.H., Seinfeld, J.H., Surratt, J.D., Szmigielski, R., Wildt, J., 2009. The formation, properties and impact of secondary organic aerosol: current and emerging issues. Atmospheric Chem. Phys. 9, 5155–5236. https://doi.org/10.5194/acp-9-5155-2009

Hanahan, D., Weinberg, R.A., 2011. Hallmarks of Cancer: The Next Generation. Cell 144, 646–674. https://doi.org/10.1016/j.cell.2011.02.013

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, Springer Series in Statistics. Springer New York, New York, NY. https://doi.org/10.1007/978-0-387-84858-7

Helma, C., Cramer, T., Kramer, S., De Raedt, L., 2004. Data Mining and Machine Learning Techniques for the Identification of Mutagenicity Inducing Substructures and Structure Activity Relationships of Noncongeneric Compounds. J. Chem. Inf. Comput. Sci. 44, 1402–1411. https://doi.org/10.1021/ci034254q

Jaganathan, K., Tayara, H., Chong, K.T., 2022. An Explainable Supervised Machine Learning Model for Predicting Respiratory Toxicity of Chemicals Using Optimal Molecular Descriptors. Pharmaceutics 14, 832. https://doi.org/10.3390/pharmaceutics14040832

Jaworska, J., Hoffmann, S., 2010. Integrated Testing Strategy (ITS) – opportunities to better use existing data and guide future testing in toxicology. ALTEX - Altern. Anim. Exp. 27, 231–242. https://doi.org/10.14573/altex.2010.4.231

Jia, X., Wang, T., Zhu, H., 2023. Advancing Computational Toxicology by Interpretable Machine Learning. Environ. Sci. Technol. 57, 17690–17706. https://doi.org/10.1021/acs.est.3c00653

Judson, R.S., Houck, K.A., Kavlock, R.J., Knudsen, T.B., Martin, M.T., Mortensen, H.M., Reif, D.M., Rotroff, D.M., Shah, I., Richard, A.M., Dix, D.J., 2010. In Vitro Screening of Environmental Chemicals for Targeted Testing Prioritization: The ToxCast Project. Environ. Health Perspect. 118, 485–492. https://doi.org/10.1289/ehp.0901392

Kang, Y., Kim, M.G., Lim, K.-M., 2023. Machine-learning based prediction models for assessing skin irritation and corrosion potential of liquid chemicals using physicochemical properties by XGBoost. Toxicol. Res. 39, 295–305. https://doi.org/10.1007/s43188-022-00168-8
Kastenhuber, E.R., Lowe, S.W., 2017. Putting p53 in Context. Cell 170, 1062–1078. https://doi.org/10.1016/j.cell.2017.08.028
Kavlock, R.J., Ankley, G., Blancato, J., Breen, M., Conolly, R., Dix, D., Houck, K., Hubal, E., Judson, R., Rabinowitz, J., Richard, A., Setzer, R.W., Shah, I., Villeneuve, D., Weber, E., 2008. Computational Toxicology—A State of the Science Mini Review. Toxicol. Sci. 103, 14–27. https://doi.org/10.1093/toxsci/kfm297
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree, in: Advances in Neural Information Processing Systems. Curran Associates, Inc.
Kelly, F.J., Fussell, J.C., 2012. Size, source and chemical composition as determinants of toxicity attributable to ambient particulate matter. Atmos. Environ. 60, 504–526. https://doi.org/10.1016/j.atmosenv.2012.06.039
Khan, F., Jaoui, M., Rudziński, K., Kwapiszewska, K., Martinez-Romero, A., Gil-Casanova, D., Lewandowski, M., Kleindienst, T.E., Offenberg, J.H., Krug, J.D., Surratt, J.D., Szmigielski, R., 2022. Cytotoxicity and oxidative stress induced by atmospheric mono-nitrophenols in human lung cells. Environ. Pollut. 301, 119010. https://doi.org/10.1016/j.envpol.2022.119010
Kim, K.-H., Kabir, E., Kabir, S., 2015. A review on the human health impact of airborne particulate matter. Environ. Int. 74, 136–143. https://doi.org/10.1016/j.envint.2014.10.005
Kind, T., Fiehn, O., 2010. Advances in structure elucidation of small molecules using mass spectrometry. Bioanal. Rev. 2, 23–60. https://doi.org/10.1007/s12566-010-0015-9
Klekota, J., Roth, F.P., 2008. Chemical substructures that enrich for biological activity. Bioinformatics 24, 2518–2525. https://doi.org/10.1093/bioinformatics/btn479
Kovacic, P., Somanathan, R., 2014. Nitroaromatic compounds: Environmental toxicity, carcinogenicity, mutagenicity, therapy and mechanism. J. Appl. Toxicol. 34, 810–824. https://doi.org/10.1002/jat.2980
Krewski, D., Acosta, D., Andersen, M., Anderson, H., Bailar, J.C., Boekelheide, K., Brent, R., Charnley, G., Cheung, V.G., Green, S., Kelsey, K.T., Kerkvliet, N.I., Li, A.A., McCray, L., Meyer, O., Patterson, R.D., Pennie, W., Scala, R.A., Solomon, G.M., Stephens, M., Yager, J., Zeise, L., 2010. TOXICITY TESTING IN THE 21ST CENTURY: A VISION AND A STRATEGY. J. Toxicol. Environ. Health B Crit. Rev. 13, 51–138. https://doi.org/10.1080/10937404.2010.483176
Kuhn, M., 2008. Building Predictive Models in R Using the caret Package. J. Stat. Softw. 28, 1–26. https://doi.org/10.18637/jss.v028.i05
Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4614-6849-3
Landrigan, P.J., Fuller, R., Acosta, N.J.R., Adeyi, O., Arnold, R., Basu, N. (Nil), Baldé, A.B., Bertollini, R., Bose-O'Reilly, S., Boufford, J.I., Breysse, P.N., Chiles, T., Mahidol, C., Coll-Seck, A.M., Cropper, M.L., Fobil, J., Fuster, V., Greenstone, M., Haines, A., Hanrahan, D., Hunter, D., Khare, M., Krupnick, A., Lanphear, B., Lohani, B., Martin, K., Mathiasen, K.V., McTeer, M.A., Murray, C.J.L., Ndahimananjara, J.D., Perera, F., Potočnik, J., Preker, A.S., Ramesh, J., Rockström, J., Salinas, C., Samson, L.D., Sandilya, K., Sly, P.D., Smith, K.R., Steiner, A., Stewart, R.B., Suk, W.A., Schayck, O.C.P. van, Yadama, G.N., Yumkella, K., Zhong, M., 2018. The Lancet Commission on pollution and health. The Lancet 391, 462–512. https://doi.org/10.1016/S0140-6736(17)32345-0
Lane, D.P., 1992. p53, guardian of the genome. Nature 358, 15–16. https://doi.org/10.1038/358015a0
Lazic, S.E., Williams, D.P., 2021. Quantifying sources of uncertainty in drug discovery predictions with probabilistic models. Artif. Intell. Life Sci. 1, 100004. https://doi.org/10.1016/j.ailsci.2021.100004

Levine, A.J., 1997. p53, the Cellular Gatekeeper for Growth and Division. Cell 88, 323–331. https://doi.org/10.1016/S0092-8674(00)81871-1
Li, R., Zhou, R., Zhang, J., 2018. Function of PM2.5 in the pathogenesis of lung cancer and chronic airway inflammatory diseases. Oncol. Lett. 15, 7506–7514. https://doi.org/10.3892/ol.2018.8355
Li, Z., Niu, F., Fan, J., Liu, Y., Rosenfeld, D., Ding, Y., 2011. Long-term impacts of aerosols on the vertical development of clouds and precipitation. Nat. Geosci. 4, 888–894. https://doi.org/10.1038/ngeo1313
Lima de Albuquerque, Y., Berger, E., Tomaz, S., George, C., Géloën, A., 2021. Evaluation of the Toxicity on Lung Cells of By-Products Present in Naphthalene Secondary Organic Aerosols. Life Basel Switz. 11, 319. https://doi.org/10.3390/life11040319
Limbu, S., Dakshanamurthy, S., 2022. Predicting Chemical Carcinogens Using a Hybrid Neural Network Deep Learning Method. Sensors 22, 8185. https://doi.org/10.3390/s22218185
Lundberg, S.M., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions, in: Advances in Neural Information Processing Systems. Curran Associates, Inc.
Mayr, A., Klambauer, G., Unterthiner, T., Hochreiter, S., 2016. DeepTox: Toxicity Prediction using Deep Learning. Front. Environ. Sci. 3. https://doi.org/10.3389/fenvs.2015.00080
Morehouse, B.R., Kumar, R.P., Matos, J.O., Olsen, S.N., Entova, S., Oprian, D.D., 2017. Functional and Structural Characterization of a (+)-Limonene Synthase from Citrus sinensis. Biochemistry 56, 1706–1715. https://doi.org/10.1021/acs.biochem.7b00143
Najjar, A., Kramer, N., Gardner, I., Hartung, T., Steger-Hartmann, T., 2023. Editorial: Advances in and applications of predictive toxicology: 2022. Front. Pharmacol. 14, 1257423. https://doi.org/10.3389/fphar.2023.1257423
Niculescu-Mizil, A., Caruana, R., 2005. Predicting good probabilities with supervised learning, in: Proceedings of the 22nd International Conference on Machine Learning - ICML '05. Presented at the the 22nd international conference, ACM Press, Bonn, Germany, pp. 625–632. https://doi.org/10.1145/1102351.1102430
OECD, 2023. (Q)SAR Assessment Framework: Guidance for the regulatory assessment of (Quantitative) Structure Activity Relationship models and predictions, OECD Series on Testing and Assessment. OECD. https://doi.org/10.1787/d96118f6-en
OECD, 2016. Testing of chemicals [WWW Document]. OECD. URL https://www.oecd.org/en/topics/testing-of-chemicals.html (accessed 8.7.25).
Olivier, M., Hollstein, M., Hainaut, P., 2010. TP53 Mutations in Human Cancers: Origins, Consequences, and Clinical Use. Cold Spring Harb. Perspect. Biol. 2, a001008. https://doi.org/10.1101/cshperspect.a001008
Palm, E., Kruve, A., 2022. Machine Learning for Absolute Quantification of Unidentified Compounds in Non-Targeted LC/HRMS. Molecules 27, 1013. https://doi.org/10.3390/molecules27031013
Peets, P., Wang, W.-C., MacLeod, M., Breitholtz, M., Martin, J.W., Kruve, A., 2022. MS2Tox Machine Learning Tool for Predicting the Ecotoxicity of Unidentified Chemicals in Water by Nontarget LC-HRMS. Environ. Sci. Technol. 56, 15508–15517. https://doi.org/10.1021/acs.est.2c02536
Pereira, K.L., Rovelli, G., Song, Y.C., Mayhew, A.W., Reid, J.P., Hamilton, J.F., 2019. A new aerosol flow reactor to study secondary organic aerosol. Atmospheric Meas. Tech. 12, 4519–4541. https://doi.org/10.5194/amt-12-4519-2019
Perwez Hussain, S., Harris, C.C., 2007. Inflammation and cancer: An ancient link with novel potentials. Int. J. Cancer 121, 2373–2380. https://doi.org/10.1002/ijc.23173
Ponce-Bobadilla, A.V., Schmitt, V., Maier, C.S., Mensing, S., Stodtmann, S., 2024. Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. Clin. Transl. Sci. 17, e70056. https://doi.org/10.1111/cts.70056

Pope III, C.A., and Dockery, D.W., 2006. Health Effects of Fine Particulate Air Pollution: Lines that Connect. J. Air Waste Manag. Assoc. 56, 709–742. https://doi.org/10.1080/10473289.2006.10464485

Pöschl, U., 2005. Atmospheric Aerosols: Composition, Transformation, Climate and Health Effects. Angew. Chem. Int. Ed. 44, 7520–7540. https://doi.org/10.1002/anie.200501122

Powers, D.M.W., 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. https://doi.org/10.48550/arXiv.2010.16061

Probst, P., Wright, M.N., Boulesteix, A.-L., 2019. Hyperparameters and tuning strategies for random forest. WIREs Data Min. Knowl. Discov. 9, e1301. https://doi.org/10.1002/widm.1301

Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D., 2022. Dataset Shift in Machine Learning. MIT Press.

Richard, A.M., Judson, R.S., Houck, K.A., Grulke, C.M., Volarath, P., Thillainadarajah, I., Yang, C., Rathman, J., Martin, M.T., Wambaugh, J.F., Knudsen, T.B., Kancherla, J., Mansouri, K., Patlewicz, G., Williams, A.J., Little, S.B., Crofton, K.M., Thomas, R.S., 2016. ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. Chem. Res. Toxicol. 29, 1225–1251. https://doi.org/10.1021/acs.chemrestox.6b00135

Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., Huang, J., 2020. Self-Supervised Graph Transformer on Large-Scale Molecular Data. https://doi.org/10.48550/arXiv.2007.02835

Russell, W.M.S., Burch, R.L., n.d. The Principles of Humane Experimental Technique.

Sandve, G.K., Nekrutenko, A., Taylor, J., Hovig, E., 2013. Ten Simple Rules for Reproducible Computational Research. PLOS Comput. Biol. 9, e1003285. https://doi.org/10.1371/journal.pcbi.1003285

Schwöbel, J.A.H., Koleva, Y.K., Enoch, S.J., Bajot, F., Hewitt, M., Madden, J.C., Roberts, D.W., Schultz, T.W., Cronin, M.T.D., 2011. Measurement and Estimation of Electrophilic Reactivity for Predictive Toxicology. Chem. Rev. 111, 2562–2596. https://doi.org/10.1021/cr100098n

Schymanski, E.L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H.P., Hollender, J., 2014. Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. Environ. Sci. Technol. 48, 2097–2098. https://doi.org/10.1021/es5002105

Seinfeld, J.H., n.d. A Wiley Interscience Publication.

Seinfeld, J.H., Pandis, S.N., 2016. Atmospheric Chemistry and Physics: From Air Pollution to Climate Change. John Wiley & Sons.

Sheridan, R.P., Wang, W.M., Liaw, A., Ma, J., Gifford, E.M., 2016. Extreme Gradient Boosting as a Method for Quantitative Structure–Activity Relationships. J. Chem. Inf. Model. 56, 2353–2360. https://doi.org/10.1021/acs.jcim.6b00591

Shrivastava, M., Cappa, C.D., Fan, J., Goldstein, A.H., Guenther, A.B., Jimenez, J.L., Kuang, C., Laskin, A., Martin, S.T., Ng, N.L., Petaja, T., Pierce, J.R., Rasch, P.J., Roldin, P., Seinfeld, J.H., Shilling, J., Smith, J.N., Thornton, J.A., Volkamer, R., Wang, J., Worsnop, D.R., Zaveri, R.A., Zelenyuk, A., Zhang, Q., 2017. Recent advances in understanding secondary organic aerosol: Implications for global climate forcing. Rev. Geophys. 55, 509–559. https://doi.org/10.1002/2016RG000540

Sobus, J.R., Wambaugh, J.F., Isaacs, K.K., Williams, A.J., McEachran, A.D., Richard, A.M., Grulke, C.M., Ulrich, E.M., Rager, J.E., Strynar, M.J., Newton, S.R., 2018. Integrating tools for non-targeted analysis research and chemical safety evaluations at the US EPA. J. Expo. Sci. Environ. Epidemiol. 28, 411–426. https://doi.org/10.1038/s41370-017-0012-y

Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., Willighagen, E., 2003. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. J. Chem. Inf. Comput. Sci. 43, 493–500. https://doi.org/10.1021/ci025584y

Sugiyama, M., Jp, C.T.A., Krauledat, M., Krauledat, M., n.d. Covariate Shift Adaptation by Importance Weighted Cross Validation.

Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P., 2003. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. J. Chem. Inf. Comput. Sci. 43, 1947–1958. https://doi.org/10.1021/ci034160g

Thangavel, P., Park, D., Lee, Y.-C., 2022. Recent Insights into Particulate Matter (PM2.5)-Mediated Toxicity in Humans: An Overview. Int. J. Environ. Res. Public. Health 19, 7511. https://doi.org/10.3390/ijerph19127511

Thomas, R.S., Bahadori, T., Buckley, T.J., Cowden, J., Deisenroth, C., Dionisio, K.L., Frithsen, J.B., Grulke, C.M., Gwinn, M.R., Harrill, J.A., Higuchi, M., Houck, K.A., Hughes, M.F., Hunter, E.S., III, Isaacs, K.K., Judson, R.S., Knudsen, T.B., Lambert, J.C., Linnenbrink, M., Martin, T.M., Newton, S.R., Padilla, S., Patlewicz, G., Paul-Friedman, K., Phillips, K.A., Richard, A.M., Sams, R., Shafer, T.J., Setzer, R.W., Shah, I., Simmons, J.E., Simmons, S.O., Singh, A., Sobus, J.R., Strynar, M., Swank, A., Tornero-Valez, R., Ulrich, E.M., Villeneuve, D.L., Wambaugh, J.F., Wetmore, B.A., Williams, A.J., 2019. The Next Generation Blueprint of Computational Toxicology at the U.S. Environmental Protection Agency. Toxicol. Sci. 169, 317–332. https://doi.org/10.1093/toxsci/kfz058

Vousden, K.H., Lane, D.P., 2007. p53 in health and disease. Nat. Rev. Mol. Cell Biol. 8, 275–283. https://doi.org/10.1038/nrm2147

Walter, M., 2022. Improving the Accuracy and Interpretability of Machine Learning Models for Toxicity Prediction (phd). University of Sheffield. https://doi.org/10/1/Walter_thesis_final_submission.pdf

Wang, Y., Xiao, J., Suzek, T.O., Zhang, J., Wang, J., Bryant, S.H., 2009. PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res. 37, W623–W633. https://doi.org/10.1093/nar/gkp456

WHO, W., 2025. Particulate matters and its health impacts [WWW Document]. URL https://www.who.int/multi-media/details/particulate-matters-and-its-health-impacts (accessed 8.18.25).

WHO, W., 2023. Ambient (outdoor) air pollution [WWW Document]. URL https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health (accessed 7.28.25).

Wickham, H., n.d. A personal history of the tidyverse.

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Bonino da Silva Santos, L., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship.

Williams, A.J., Grulke, C.M., Edwards, J., McEachran, A.D., Mansouri, K., Baker, N.C., Patlewicz, G., Shah, I., Wambaugh, J.F., Judson, R.S., Richard, A.M., 2017. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. J. Cheminformatics 9, 61. https://doi.org/10.1186/s13321-017-0247-6

Witkowski, B., Gierczak, T., 2017. Characterization of the limonene oxidation products with liquid chromatography coupled to the tandem mass spectrometry. Atmos. Environ. 154, 297–307. https://doi.org/10.1016/j.atmosenv.2017.02.005

Wu, X., Zhou, Q., Mu, L., Hu, X., 2022. Machine learning in the identification, prediction and exploration of environmental toxicology: Challenges and perspectives. J. Hazard. Mater. 438, 129487. https://doi.org/10.1016/j.jhazmat.2022.129487

Youden, W.J., 1950. Index for rating diagnostic tests. Cancer 3, 32–35. https://doi.org/10.1002/1097-0142(1950)3:1%253C32::AID-CNCR2820030106%253E3.0.CO;2-3

Yvan, S., 2007. (PDF) A review of feature selection techniques in bioinformatics. ResearchGate. https://doi.org/10.1093/bioinformatics/btm344

Zeiger, E., 2019. The test that changed the world: The Ames test and the regulation of chemicals. Mutat. Res. Toxicol. Environ. Mutagen. 841, 43–48. https://doi.org/10.1016/j.mrgentox.2019.05.007

Zeron-Medina, J., Wang, X., Repapi, E., Campbell, M.R., Su, D., Castro-Giner, F., Davies, B., Peterse, E.F.P., Sacilotto, N., Walker, G.J., Terzian, T., Tomlinson, I.P., Box, N.F., Meinshausen, N., De Val, S., Bell, D.A., Bond, G.L., 2013. A Polymorphic p53 Response Element in KIT Ligand Influences Cancer Risk and Has Undergone Natural Selection. Cell 155, 410–422. https://doi.org/10.1016/j.cell.2013.09.017

Appendix

Table1: Training results for XGBoost models with different hyperparameters (nrounds = 100, 200, 300).

Metric	nrounds = 100	nrounds = 200	nrounds = 300
Best max_depth	6	9	9
Best eta	0.01	0.01	0.01
Best gamma	5	5	5
Colsample_bytree	0.6	0.6	0.6
min_child_weight	1	1	1
subsample	0.6	0.6	0.6
Accuracy	79.02%	64.34%	66.43%
Sensitivity (Recall)	66.67%	87.50%	83.33%
Specificity	81.51%	59.66%	63.03%
Precision (PPV)	42.11%	30.43%	31.25%
Negative Predictive Value (NPV)	92.38%	95.95%	94.94%
Balanced Accuracy	74.09%	73.58%	73.18%
Kappa	0.3908	0.2698	0.2784
True Positives (TP)	16	21	20
False Positives (FP)	22	48	44
False Negatives (FN)	8	3	4
True Negatives (TN)	97	71	75

I chose nrounds = 100 because, for this prediction task, it provided the strongest overall training metrics by best balancing sensitivity correctly flagging carcinogens and specificity preventing false alarms about harmless chemicals while capturing signal without overfitting (unlike 200–300 rounds, which increased false positives).

Table2: Final XGBoost (trained on full data)

Section	Metric	Value
Hyperparameters	nrounds	100
	max_depth	6
	eta	0.01
	gamma	5
	colsample_bytree	0.60

	min_child_weight	1
	subsample	0.60
Performance	Optimal ROC threshold (Youden's J)	0.4805
	Accuracy	90.25%
	95% CI for Accuracy	87.84% – 92.32%
	Kappa	0.7131
	Sensitivity (Recall)	93.70%
	Specificity	89.51%
	Precision (PPV)	65.75%
	Negative Predictive Value (NPV)	98.51%
	Balanced Accuracy	91.61%
	Prevalence	17.69%
	Detection Rate	16.57%
	Detection Prevalence	25.21%
	No Information Rate	82.31%
	P-value [Acc > NIR]	1.562×10^{-9}
	McNemar's Test p-value	2.378×10^{-10}
Confusion matrix	True Positives (TP)	119
	False Positives (FP)	62
	False Negatives (FN)	8
	True Negatives (TN)	529
	Total N	718

Table3: Explanations of XGBoost Parameters

Parameter	**Plain meaning**	**What higher values do**	**What lower values do**	**Typical range**
nrounds	Number of boosting iterations (trees).	More trees - can learn more complex patterns, but risk overfitting and longer training time.	Fewer trees - faster, lower variance, but risk underfitting.	50–1000
max_depth	Maximum tree depth (complexity per tree).	Deeper trees capture interactions but can overfit and reduce generalisation.	Shallower trees are more regularised, may miss	3–12

			complex signals.	
eta (learning rate)	Step size for each boosting update.	Larger steps learn faster but can overshoot/overfit.	Smaller steps learn slowly, more stable, need more trees.	0.01–0.3
gamma	Minimum loss reduction to split a node (split penalty).	Higher gamma prunes weak splits - simpler, less overfit.	Lower gamma allows more splits - finer fit, risk overfit.	0–10
colsample_bytree	Fraction of features sampled per tree.	Lower fraction adds randomness - less overfit, may reduce accuracy.	Higher fraction uses more features - can increase fit but overfit risk.	0.3–1.0
min_child_weight	Minimum sum of Hessian (roughly, min data weight) in a leaf.	Higher - require more data per leaf - smoother, less overfit.	Lower - smaller leaves allowed - finer fits, overfit risk.	1–10
subsample	Fraction of rows sampled per tree.	Lower adds randomness - regularises; too low can underfit.	Higher uses more data per tree - stronger fit, overfit risk.	0.5–1.0