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

**Abstract**

Natural and anthropogenic aerosol atmospheric aerosols have the potential to carry carcinogenic chemicals that can harm public health. Conventional tests for toxicity, although well established, are time-consuming, expensive, and ethically limited 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.

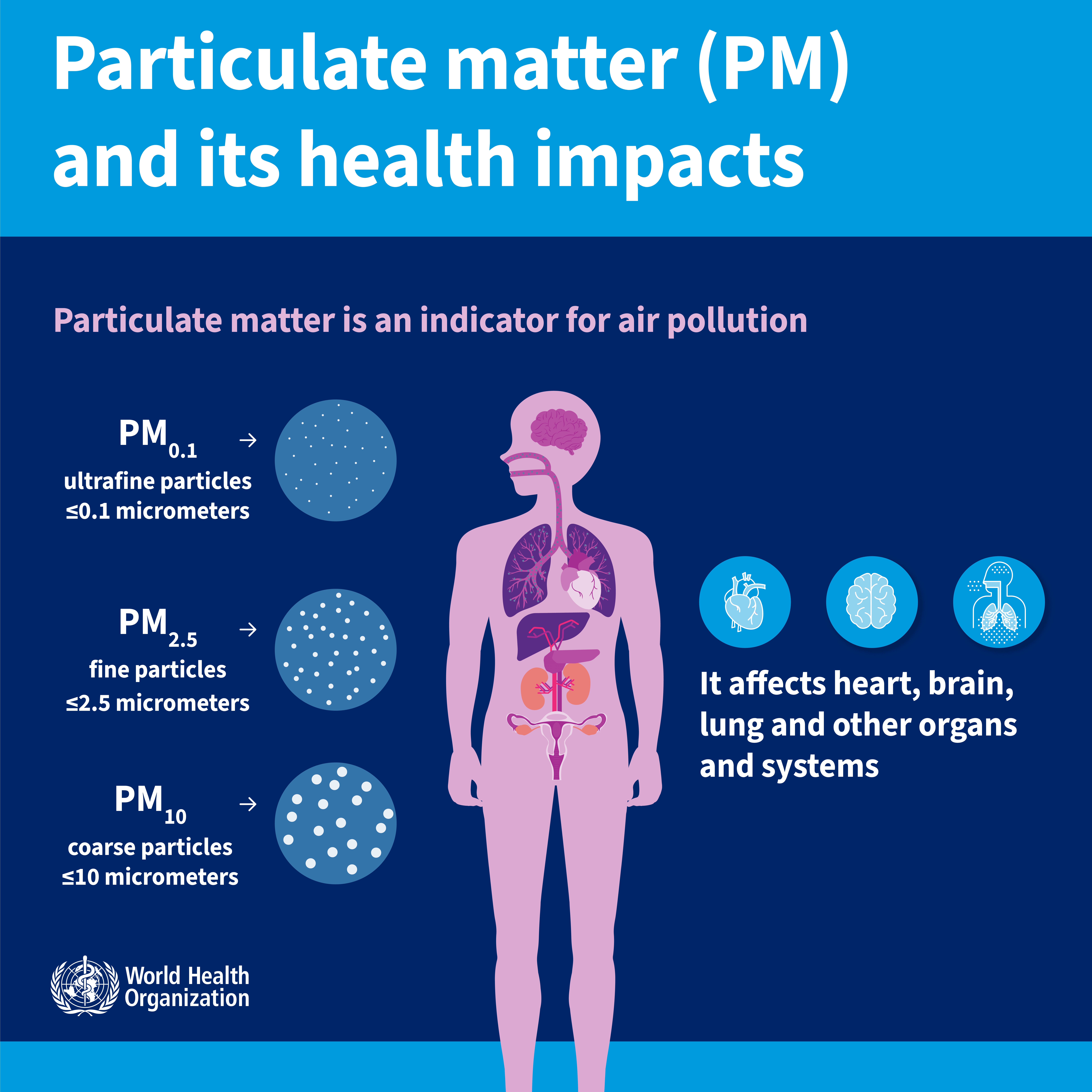
Data were extensively pre-processed for chemical quality and relevance before model training. Validation was initially performed with LC-MS spectra from MassBank, which were processed in SIRIUS software to produce structure-based predictions for direct comparison of model output. The model was then transferred 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.

**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 (PM₂.₅) 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 PM₂.₅ particle penetration into the respiratory tract and bloodstream, highlighting pathways of systemic health effects (WHO, 2025).

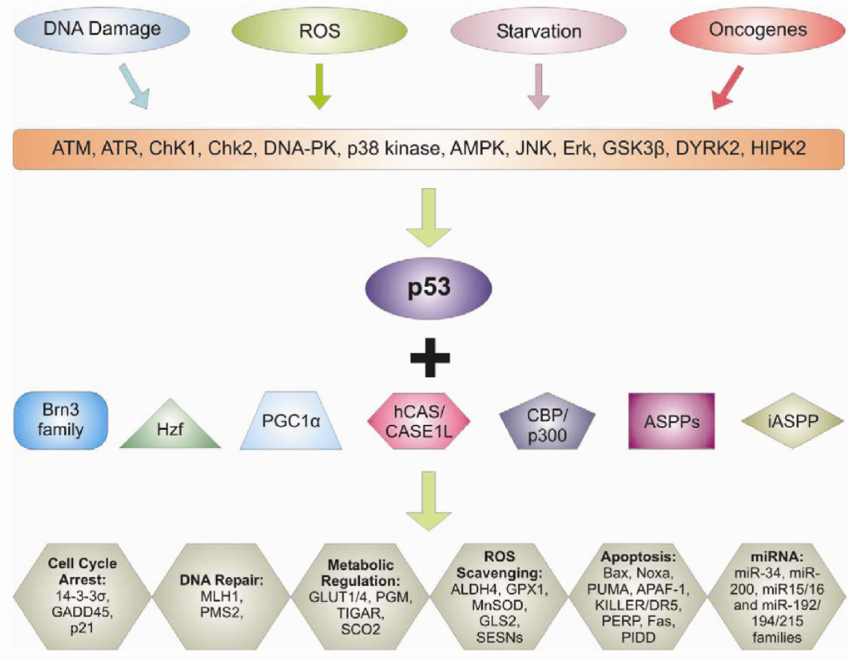
The danger posed by PM₂.₅ is not just because of the infinitesimal size but also because of the compound multivalency of the chemical structure of PM₂.₅. These particles have a propensity to transfer a mix of the toxicants of the polycyclic aromatic hydrocarbons (PAHs) and the volatile organic compounds (VOCs) to 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, PM₂.₅ 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 with thorough, 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 the progressive accumulation of genetic mutations that inactivate the body's control systems so cells can grow uncontrollably, evade immune defences, and survive after DNA damage (Hanahan and Weinberg, 2011). At the core of the body's defence systems is the p53 protein a tumour suppressor that is tasked with having a central role to detect and respond to genomic instability.

In normal conditions, p53 lies in a dormant state. But with cell or DNA injury, p53 is quickly induced to halt the cell cycle, initiate DNA repair, or trigger apoptosis, which averts the growth of damaged cells (Lane, 1992), (Levine, n.d.). TP53 gene encoding p53 or in its regulatory network mutations are among the most common alterations in human cancers, occurring in more than half of all cases (Olivier et al., 2010), (Kastenhuber and Lowe, 2017).



**Figure 1.2.** Simplified schematic of the p53 tumour suppressor pathway, showing its role in DNA repair, cell cycle arrest, and apoptosis.

Environmental carcinogens like high doses of chemicals in contaminated air can interfere with p53-mediated pathways by either inducing DNA damage or by physically disrupting the activity of p53 (Perwez Hussain and Harris, 2007), (Zeron-Medina et al., 2013). For this reason, the p53 pathway is thought to be an effective mechanistic target for the identification of potentially carcinogenic chemicals.

Technologies such as the ATG\_p53\_CIS assay with high throughput, designed through a collaborative effort between the U.S. Environmental Protection Agency and the National Center for Advancing Translational Sciences, offer an effective way of screening large libraries of chemicals. The assay uses human cells that contain a fluorescent reporter system that is designed to detect alterations in the p53 pathway, assigning a simple "active" or "inactive" label (Judson et al., 2010), (Richard et al., 2016). Such methods provide a scalable method of mechanistic toxicity evaluation, enabling early detection of environmental hazards before they spread far and wide.

**1.3 Traditional and Computational Approaches to Chemical Toxicity Evaluation**

Animal testing dominates conventional toxicology, particularly the application of rodent bioassays that take months or years and examine cancer development after chemical exposure, (Zeiger, 2019), (Krewski et al., 2010). While the tests are useful in yielding data, they take time, are expensive, and pose ethical concerns due to extensive use of animals. Further, species-specificity reduces their applicability to human health (Basketter et al., 2012).

One of the drawbacks is the limited scope: animal tests only expose single chemicals isolated, but human beings get exposed to mixtures that are complex. The CompTox Chemicals Dashboard contains over 875,000 entries, but only the majority have limited toxicological profiles (Williams et al., 2017). New approach methodologies (NAMs) are, therefore, gaining greater attention for their ability to offer faster, mechanistically relevant, and animal-free alternatives (OECD, 2016).

High-throughput in vitro tests, organ-on-a-chip devices, and computer models for forecasting molecular structure-based toxicity and bioactivity are a few of the novel methods. The novel

approaches have been applied to programs such as ToxCast and Tox21 for thousands of chemicals on many biological pathways, e.g., carcinogenesis (Dix et al., 2007), (Kavlock et al., 2008).

Of computational methods, logistic regression, random forests, and support vector machines (SVMs) were early models used. Each is strong logistic regression is easy to interpret; SVMs are able to handle difficult boundaries; and random forests reduce variance but all three struggle with scale and interpretability in large chemical datasets.

Over the last few years, advancements have incorporated deep learning and ensemble methods such as XGBoost, which have an improved performance, especially in high-dimensional settings. Graph neural networks (GNNs) represent an emerging field that is capable of learning directly from molecular graphs, although only just starting to be applied to toxicology. Interpretation techniques such as SHAP (Shapley Additive Explanations) have also enabled improved transparency through the attribution of importance scores to individual features (Lundberg and Lee, 2017).

Together, they expand the toxicologist's toolkit, enabling faster and larger-scale assessment of chemical hazard.

**1.4 Data Science Meets Toxicology: In Silico Toxicity Prediction of Chemicals by 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. Such an approach, referred to as in silico toxicology, presents a cost-saving and scalable alternative to traditional experiments (Cavasotto and Scardino, 2022; Wu et al., 2022).

The popular gradient boosting algorithm XGBoost is particularly well-suited for such an endeavor because it is efficient, supports missing values, and is immune to class imbalance (Chen and Guestrin, 2016; Kang et al., 2023). Paired with interpretability methods such as SHAP, XGBoost models offer interpretability through tracing the exact chemical properties that affect the learning of toxicity (Lundberg and Lee, 2017; Ponce-Bobadilla et al., 2024).

Several studies demonstrate the growing power of ML in toxicology. Cornell applied XGBoost to predict carcinogenicity of aerosols from LC–MS, whereas (Peets et al., 2022) developed MS2Tox to predict ecotoxicity from MS2 fragmentations. (Dührkop et al., 2019) outlined the CSI:FingerID and SIRIUS frameworks for structure and activity forecasting from MS/MS spectra, enabling toxicity prediction even in the absence of full molecular identities. (Arturi and Hollender, 2023)visualised risk-driven prioritisation of new environmental chemicals using Random Forests, while (Jaganathan et al., 2022) developed an interpretable model of respiratory toxicity through optimal molecular descriptors. Similarly, (Jia et al., 2023) highlighted how explainable ML can benefit computational toxicology by strengthening regulatory trust in predictions.

Apart from endpoints of toxicity, (Palm and Kruve, 2022) illustrated that ML can facilitate absolute quantification of unidentified compounds in LC/HRMS processes, while (Born et al., 2023)integrated machine learning into digital discovery platforms to enhance chemical investigation. More advanced methods such as Graph Neural Networks (GNNs) are currently under investigation as useful tools that can learn from molecular graphs directly, with new potential in predicting toxicity (Rong et al., 2020).

Together, these studies illustrate how ML methods spanning from classical ensemble models to deep learning and GNNs are bridging gaps in classical toxicology through data-driven predictions to support regulatory screening and hypothesis generation.

**1.5 Bridging Prediction and Reality: Validation Using Mass Spectrometry Data**

Models do not matter if they run under idealized conditions. In environmental toxicology, model predictions have to be validated against compounds encountered in the atmosphere with routine non-targeted analysis by means like LC-MS/MS (Kind and Fiehn, 2010).

LC-MS/MS is a method by which researchers are able to distinguish, detect, and characterise thousands of air sample compounds. SOAs that significantly contribute to PM2.5 can be profiled through this method (Shrivastava et al., 2017), (Witkowski and Gierczak, 2017). The majority of the compounds discovered are uncharacterised and do not exist in regulatory databases.

To interpret these spectra, programs like SIRIUS and CSI:FingerID generate molecular formulas and predict substructural fingerprints from fragmentation patterns (Dührkop et al., 2019, 2015). The predicted fingerprints may then be provided to ML models trained on knowns to predict toxicity in unknowns.

(Peets et al., 2022) demonstrated this method with XGBoost and MS2Tox for predicting toxicity from spectral fingerprints. The approach generalizes toxicity prediction to as yet uncatalogued chemicals, allowing real-time chemical surveillance and prioritisation.

With the same method here: SIRIUS is applied to treat spectral data from the LCSB MassBank, and predicted fingerprints are used as input to the trained XGBoost model to predict possible activation of the p53 pathway. This allows for validation against real-world environmental exposures instead of curated data sets.

**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 unsound (Zeiger, 2019), (Basketter et al., 2012). newer methods, such as high-throughput screening (HTS) and computer toxicology, can screen more rapidly, in volume, and with minimal reference to animal studies (OECD, 2016), (Jaworska and Hoffmann, 2010).

The goal of this research is to build a machine learning model that can forecast the carcinogenicity of atmospheric chemicals based on molecular fingerprints and p53 assay data. The model presents a faster, scalable alternative to traditional toxicology methods. A machine learning model known as XGBoost, is employed within this research to predict whether a chemical has the potential to interfere with a cancer-related biological pathway, based on its molecular "fingerprints" and high-throughput testing data. Molecular fingerprints are numeric labels that capture the structure of a molecule in a form that can be processed by a computer (Guha, 2007). High-throughput assays are automated laboratory experiments that assay quickly the number of chemicals that interact with a given biological process (Kavlock et al., 2008).

To validate the model's viability for practical use, it is also tested against analytical data from environmental samples analyzed through tandem mass spectrometry (MS/MS). Predictions of chemical structures from patterns in spectra are made through computational tools like CSI:FingerID , similar to pollutants in air monitoring studies.

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 so that scientists and regulators can better pinpoint high-risk air pollutants with greater speed, particularly in industrially or traffic-exposed locations (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, 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, n.d.).

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 on whether they are able to activate a signal in the ATG\_P53\_CIS assay. Only organic compounds with good 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**

To convert chemical structures into computational inputs, a wide variety of molecular fingerprints were computed using the rcdk package in R (Guha, 2007). Fingerprints are numerical or binary descriptors that capture the presence or absence of substructural motifs, topological descriptors, and physicochemical descriptors.

A range of fingerprints were used in this study to offer a high density and diversified coverage of structure data. These included MACCS keys (166-bit substructure patterns), PubChem fingerprints (881-bit features utilized by the PubChem database), and Klekota-Roth fingerprints (4860-bit patterns optimized for bioactivity prediction). Electronic and topological properties were also extracted by using substructure-based fingerprints such as the CDK standard (1024-bit) and EState fingerprints.

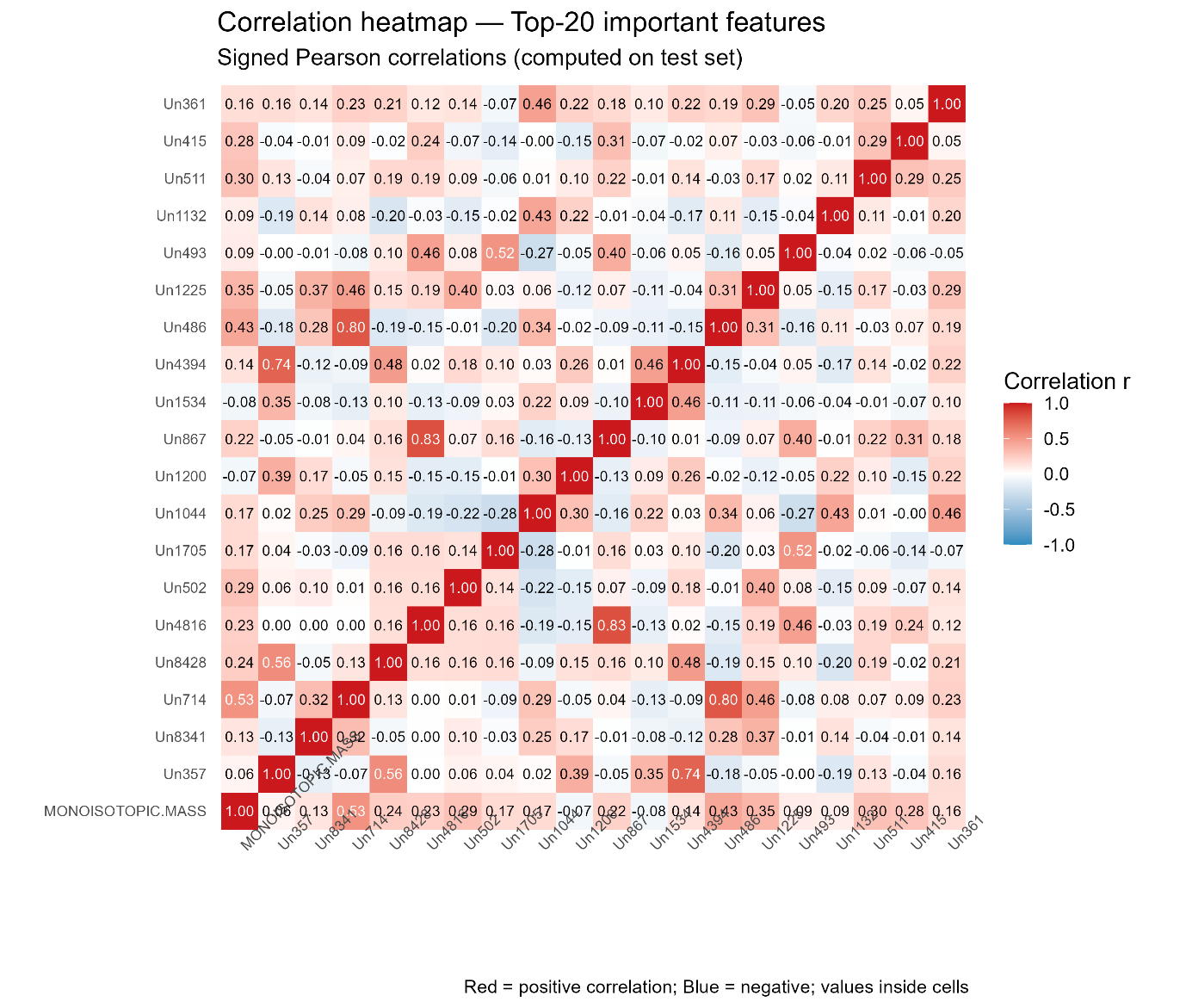
In addition to these canonical descriptors, particular SMARTS-based fingerprints were designed with particular reactive substructures pertinent to carcinogenicity. In addition to this, ring system counts, atom-type distributions, and monoisotopic mass were included in the resulting feature matrix. This comprehensive approach ensured that the model was given both general and particular chemistry information that may be pertinent to carcinogenic activity.

All the fingerprint vectors were normalized and projected to retain only common features between compounds. Near-zero variance features were removed to reduce noise, and highly correlated features (correlation > 0.9) were removed to prevent multicollinearity. Such preprocessing is very important in cheminformatics to make the model stable and generalizable (Svetnik et al., 2003).

**2.3 Data Preprocessing and Feature Selection**

Preprocessing of data was performed with the caret and dplyr packages. Missing values were removed after computing all cheminformatics features and a full cases dataset was retained. Monoisotopic mass was z-score normalized to scale all variables to a common scale.

Near-zero variance columns were identified with the nearZeroVar function of the caret package. These columns contribute minimally to the variation and can be detrimental to the efficiency of models, especially to tree-based models. Correlated columns were eliminated using Pearson correlation matrices in order to reduce redundancy and overfitting.



**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 class outcome variable HIT.CALL was also converted to a numeric vector where 1 represents actives and 0 represents inactives. In class imbalance, the ratio of inactives to actives was used to make an estimate of a scale\_pos\_weight parameter for the XGBoost classifier. Class imbalance exists in toxicology data sets, and weighting schemes avoid biasing against the majority class (Chawla et al., 2002).

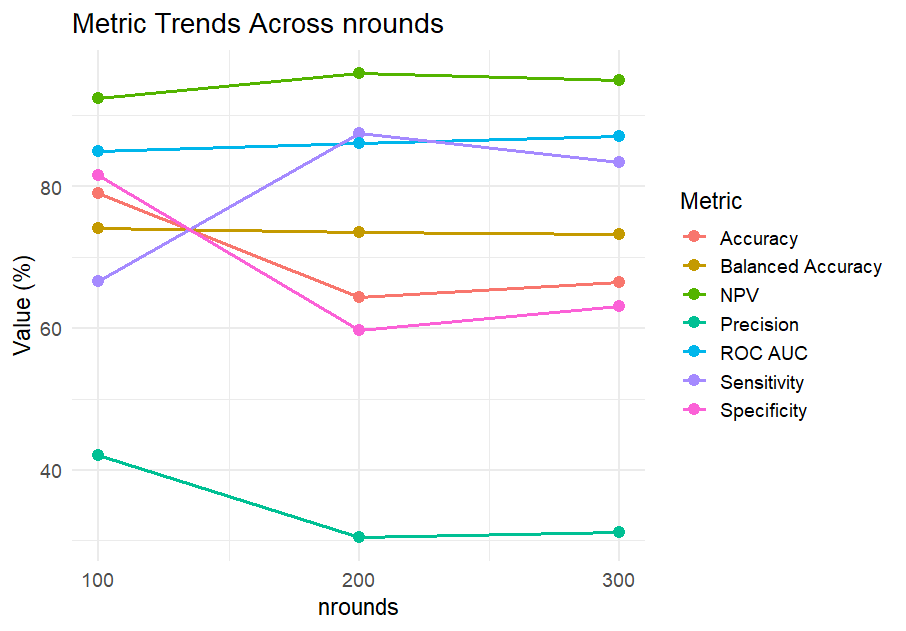
Train/test splits were carried out with an 80:20 stratified split in order to keep the actives proportional in the sets. Such subsets were used while model tuning and cross-validation was done in order to test for generalisability. Data matrices were converted to sparse matrices by using xgb.DMatrix so that computational efficiency is preserved.

**2.4 XGBoost Machine Learning Model Development**

For the prediction of carcinogenic activity from structural fingerprints, the modelling framework was the XGBoost (Extreme Gradient Boosting) gradient boosting algorithm. XGBoost is an ensemble learning algorithm in which it builds a sequence of decision trees where each successive tree seeks to reduce the error of predictions made by the earlier ones (Chen and Guestrin, 2016). This method is particularly well-suited to high-dimensional, skew data sets like those encountered in cheminformatics, as it is very scalable, can regularise, and does well for all classification tasks (Mayr et al., 2016; Sheridan et al., 2016)

XGBoost optimizes a differentiable loss function using second-order approximation for enhanced convergence. It employs L1 and L2 regularisation, column and row subsampling, and shrinking learning rate for reducing overfitting. It applies parameters such as gamma (threshold of loss reduction before another partition) and eta (learning rate) to manage model generalisation and complexity. Parameters colsample\_bytree and subsample\_bytree restrict overreliance on a particular feature or sample, enhancing model stability (Probst et al., 2019; Walter, 2022).

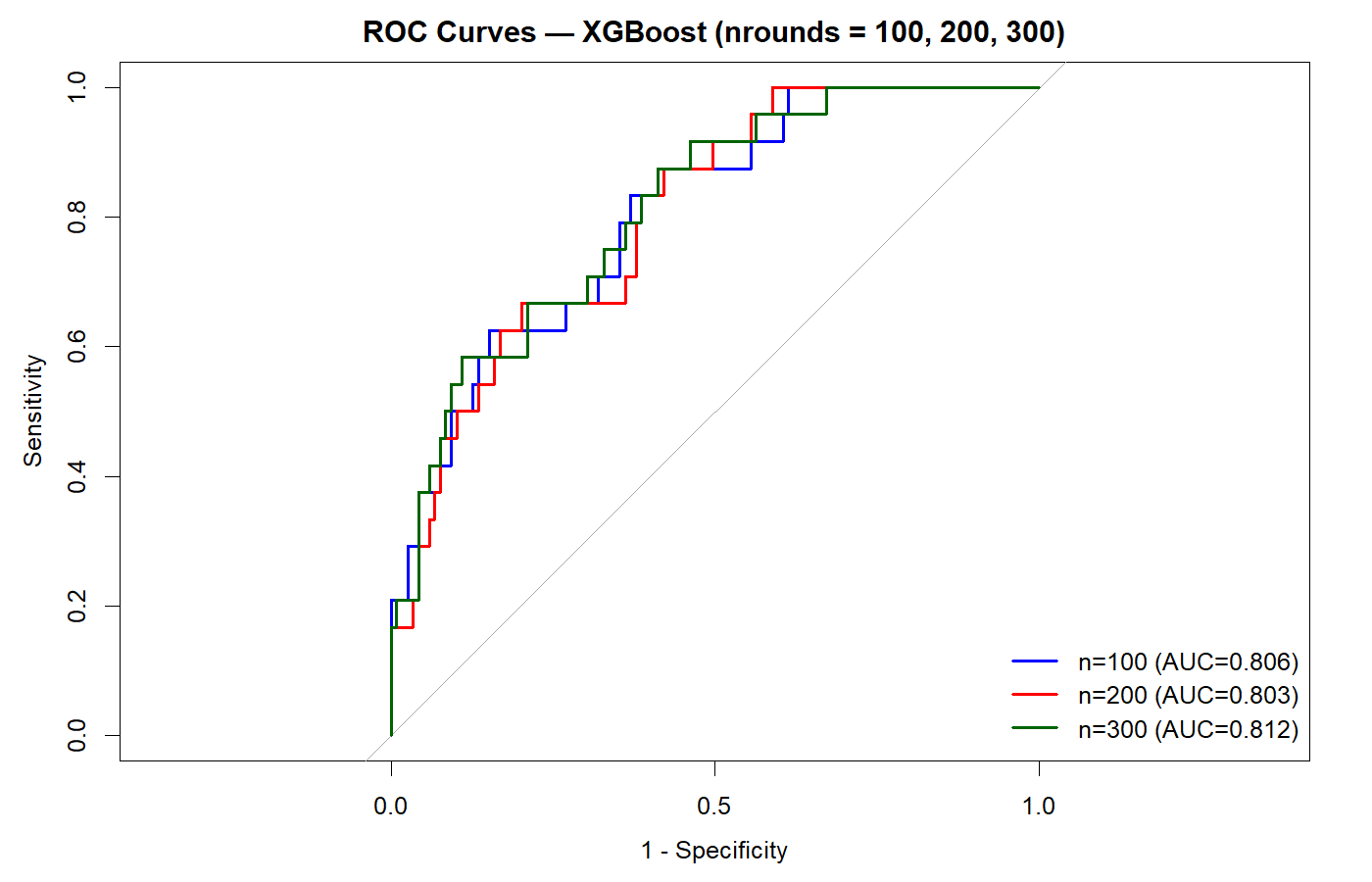
Model construction initially incorporated 10-fold cross-validation with caret in R and a hyperparameter grid. The setup was, however, slow and did not lead to enhanced precision. This was addressed by altering the workflow to use 5-fold cross-validation, which is generally sufficient in toxicology modelling and significantly reduces runtime (Boulesteix et al., 2017). Rather than grid-searching all parameters at once, the models were each trained separately on individual sets of nrounds (boosting rounds) so as to facilitate comparative evaluation. The modular design was easier to comprehend and computationally efficient (Hutter et al., 2019).



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

The best-performing model on the training set used nrounds = 100, eta = 0.01, max\_depth = 6, gamma = 5, colsample\_bytree = 0.6, and subsample = 0.6. After this setup had been chosen, the final model was then retrained on the entire data set (718 compounds) so that the learning algorithm could make use of all the structural and toxicity data available. This is a process endorsed in machine learning literature for small to medium-sized data sets where generalisability to hold-out validation is more critical (Kuhn and Johnson, 2013).

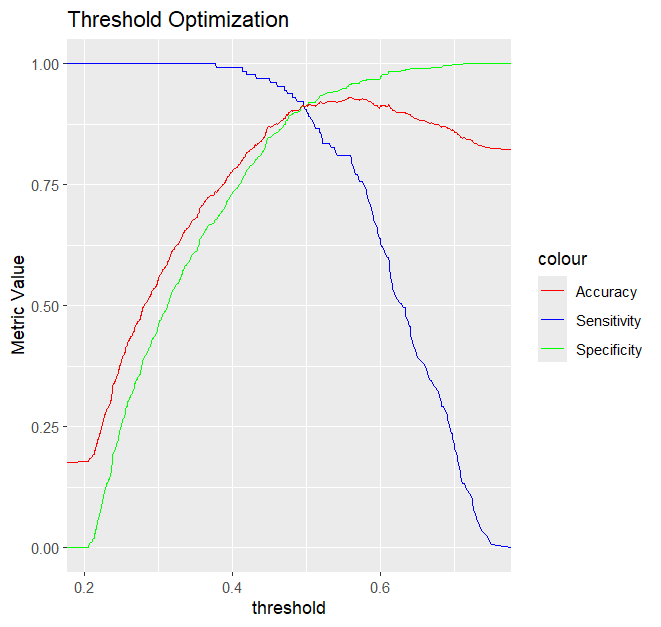
Instead of relying on the default classification threshold (0.5), the best cutoff from the ROC curve was determined by the study using Youden's J statistic. The use is generally recognized in medical diagnosis as well as binary classification issues based on its capacity to trade-off sensitivity and specificity (Powers, 2020; Youden, 1950).



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

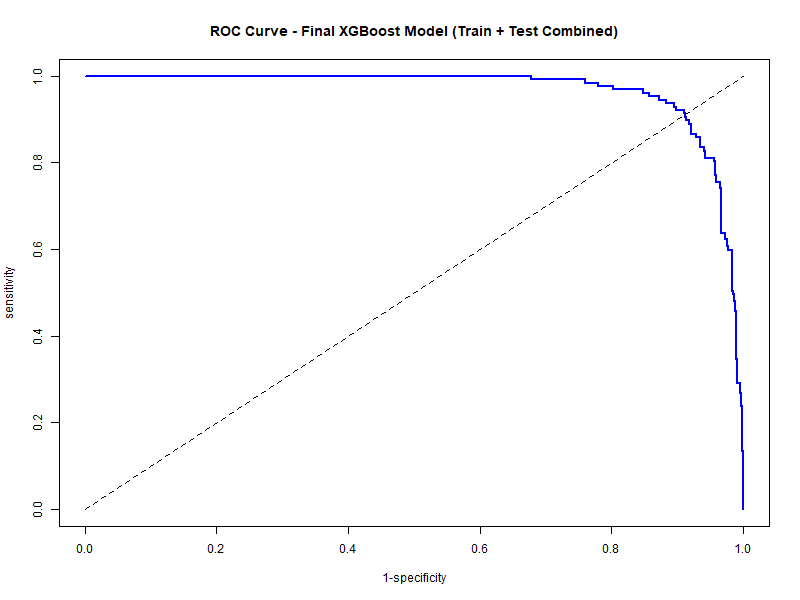
Two thresholds were used to illustrate the impact on classification performance. At decision threshold = 0.47, the model optimized for sensitivity and selected 118 out of 127 true positives and achieved a sensitivity of 92.91% and balanced accuracy of 82.58%. However, this was at a price of 164 false positives and reduced precision of 41.84%, indicating the compromise between recall and specificity in the conventional way. This setting is especially excellent in those situations where missing a positive instance is riskier, for example, in cancer or toxicological risk detection (Chawla et al., 2002; Fernandez-Delgado et al., n.d.).

Using the more traditional value of 0.65, the false positives were reduced to 40 with significantly greater specificity of 93.23% and precision of 65.22%. This came at the cost of sensitivity, falling to 59.06%, detecting only 75 of the 127 active compounds. The overall accuracy increased to 87.19%, showing that such a parameter setting for the threshold is preferable for the case where false positives are costly, i.e., regulatory decision-making or high-priority chemical screening (Hastie et al., 2009).



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

Following threshold optimisation, the last XGBoost model was retrained on the entire data with optimised hyperparameters. The best ROC-derived threshold in the retrained model was 0.4805. The model achieved 90.25% accuracy, 93.70% sensitivity, 89.51% specificity, and 91.61% balanced accuracy at this threshold. The Kappa statistic was 0.7131, indicating high agreement between predicted and actual labels greater than by chance (Brodersen et al., 2010; Cohen, 1960). These findings confirm that the model had generalized well in active as well as inactive classes.



**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.

Performance testing for all was done using pROC and caret packages in R. Confusion matrices were built, and the performance was represented using ROC curves and performance metrics. Reproducibility of the model was maintained with fixed random seeds usage and saveRDS usage in saving trained model objects. Such practice follows computational modelling and reproducibility best practices (Sandve et al., 2013; Wilkinson et al., 2016).

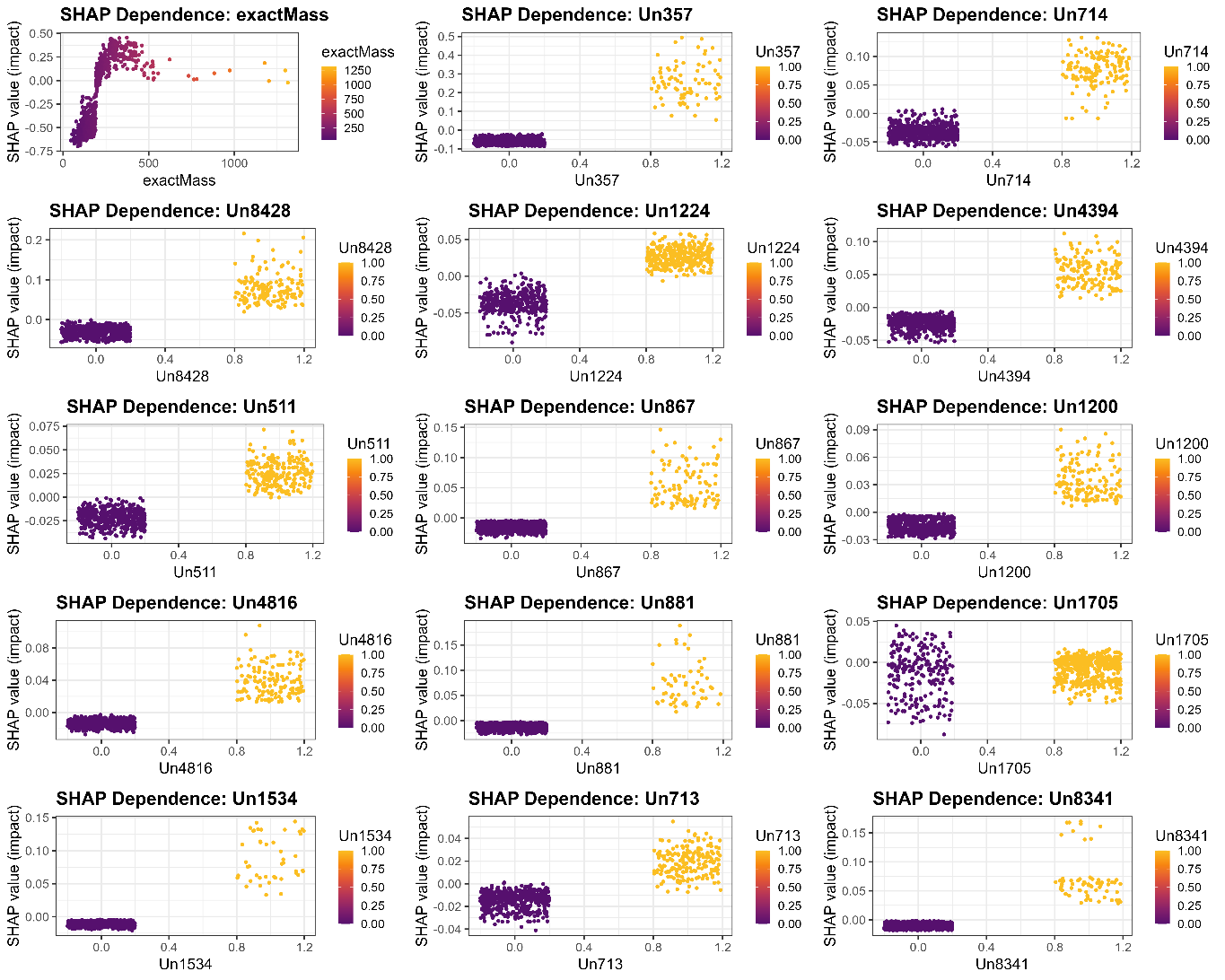
Every modelling decision, from parameter optimisation, choice of evaluation metric to threshold optimisation, was cheminformatics and toxicology data science best practice-informed. Careful design of training stages, application of ROC-directed optimisation, and retraining with combined data intended to result in a model effective as well as capable of becoming flexible enough to accommodate multiple real-world scenarios of chemical safety assessment.

**2.5 External Validation Using Mass Spectral Data**

Without external validation, machine learning models are likely to be limited to idealized datasets, which consequently limits their transferability to real monitoring. The trained XGBoost classifier was therefore validated using high-resolution tandem mass spectrometry (MS/MS) data aggregated from the Luxembourg Centre for Systems Biomedicine (LCSB) MassBank, which is heavily utilized in non-target screening in the absence of complete structure (Kind and Fiehn, 2010; Schymanski et al., 2014).

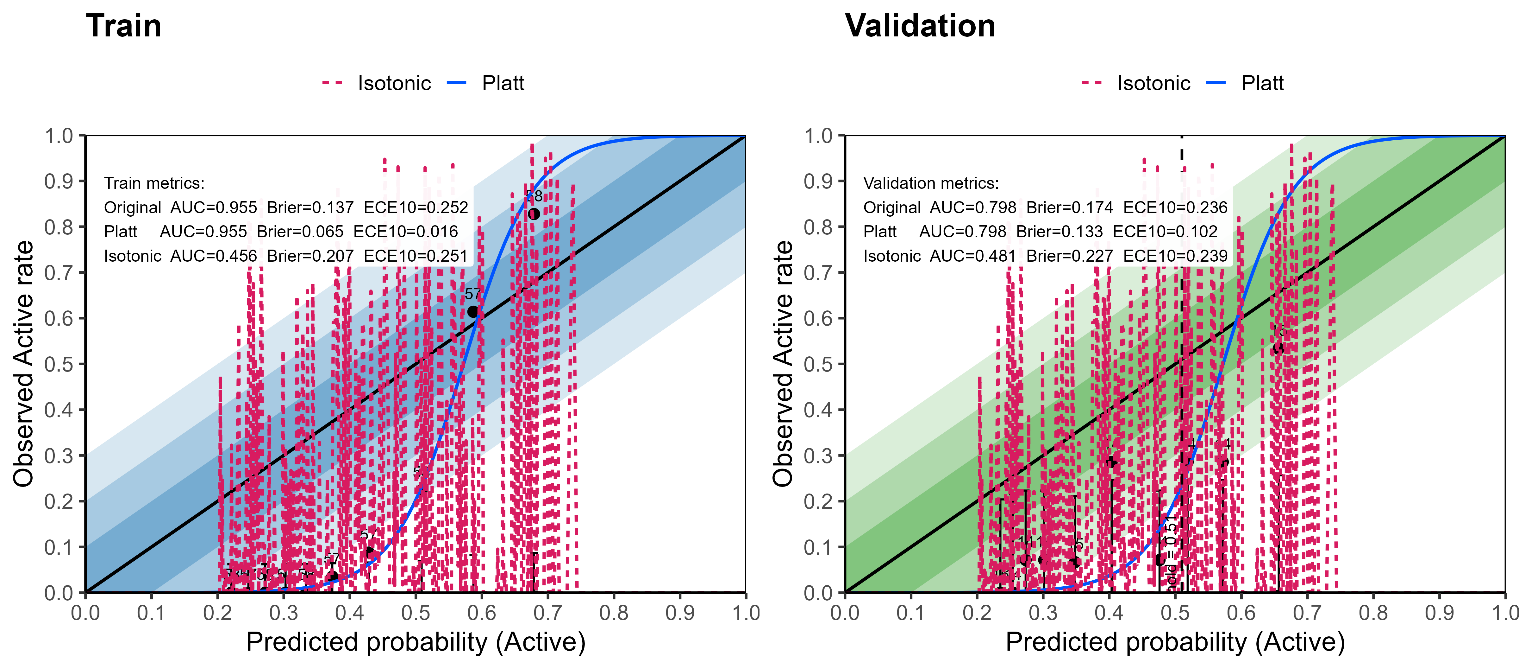
The first dataset contained over 5,500 compounds. After filtering for valid identifiers, complete metadata, and experimentally validated structures, the entries were matched against toxicology databases to provide p53 activity labels (Judson et al., 2010; Richard et al., 2016). Since non-target workflows preferentially classify compounds at the elemental formula level, harmonisation and de-duplication was performed at that level, leaving a final validation set of 319 compounds with high-confidence annotations (Schymanski et al., 2014).

Predictive MS/MS fragmentation-based molecular fingerprints for the compounds were calculated from MS/MS fragmentations using CSI:FingerID in SIRIUS v5.6.3 (Dührkop et al., 2019). Predictive fingerprints are good approximations of structural descriptors used for model training but directly computable from spectra, showing a well-documented "domain shift" between training and test representations (Quinonero-Candela et al., 2022; Sugiyama et al., n.d.). After alignment of features to development schema and same preprocessing procedures, the model was applied in making predictions based on the optimised threshold established during development. Use of SHAP in toxicology modelling has been emphasised as critical for interpretability (Walter, 2022)



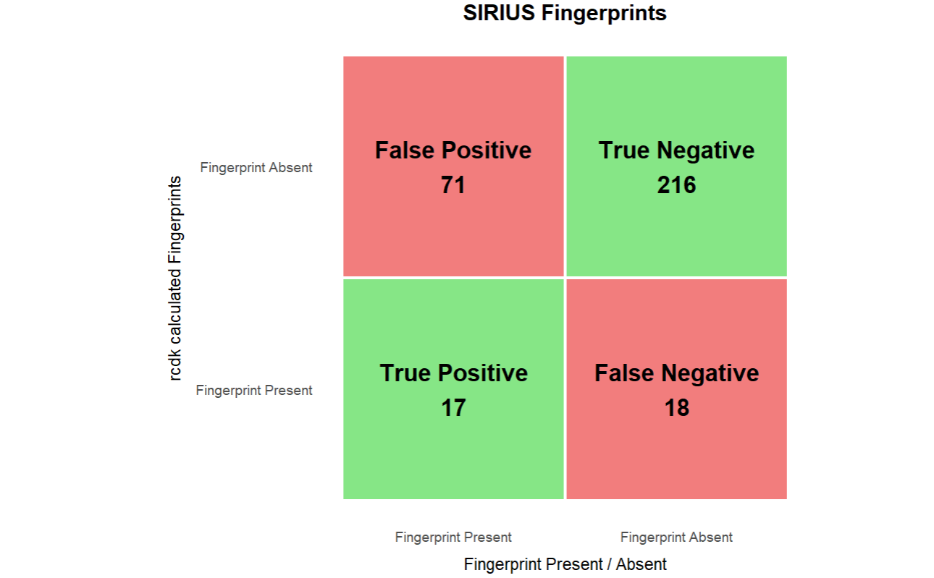
**Figure 2.5.1.** SHAP dependence plots for the top 15 features (e.g., exact mass, Un714, Un8428), showing how increasing or decreasing feature values (colour-coded by intensity) influence carcinogenicity predictions.

To determine whether probability predictions were reliable, calibration was examined. Well-calibrated models produce probabilities that match observed frequencies, a factor in risk assessment (Guo et al., n.d.; Niculescu-Mizil and Caruana, 2005) Figure 2.5.2 shows calibration plots for Platt scaling and isotonic regression comparing training and validation sets. Platt scaling, which calibrates the raw model outputs through a logistic curve, led to lower Brier scores (a measure of how close predicted probabilities are to the actual outcomes) and expected calibration error (ECE), the discrepancy between predicted probabilities and empirical frequencies, i.e., it produced more trustworthy probability estimates, while isotonic regression, a step-wise fitting method, overfit.

****

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

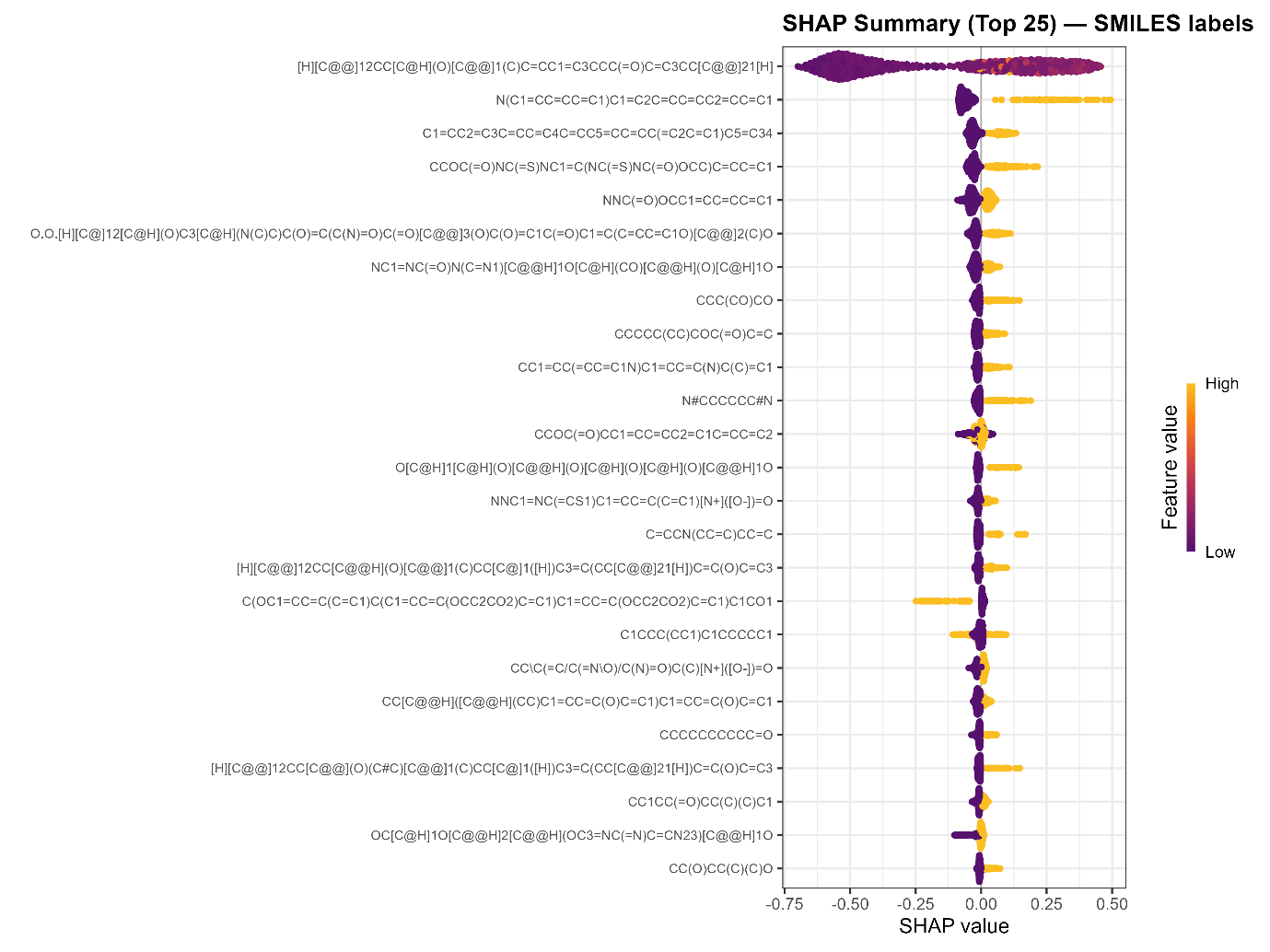
On this extrinsic dataset, the model worked at a global accuracy of 72.36%, as evidenced by the confusion matrix (Figure 2.5.3). The contingency table using the matrix shows the true positives, false positives, true negatives, and false negatives, a clear representation of prediction outcome versus reference assay label. This type of behaviour is a manifestation of correct classification of inactive compounds as well as caution in labelling as positives, something that has been mimicked by previous studies using models across different descriptor modalities (Dührkop et al., 2019; Peets et al., 2022).



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

Accuracy thus attained assures that the model remains predictive even when executed on spectrum-derived descriptors rather than directly on known structures. This is an important step towards environmental monitoring relevance, where many atmospheric or industrial chemicals are identified via high-resolution mass spectrometry but not completely characterised (Arturi and Hollender, 2023). At the same time, though, it must be realized that forecast is never absolute but rather probable. ~72% accuracy is both representative of the model's strength as well as weakness since part of the error lies in inherent information loss when fingerprints are derived from spectra rather than being derived from full molecular graphs.

Through demonstration of predictions to also extend to analytically characterized but structurally elusive compounds, this validation establishes the potential of the framework for application in real-world chemical prioritisation and regulatory toxicology. It is also a follow-through of novel principles of sustainable toxicology through combining computation-based prediction with high-throughput analytical chemistry to reduce the use of animals while expanding coverage to untested chemicals.



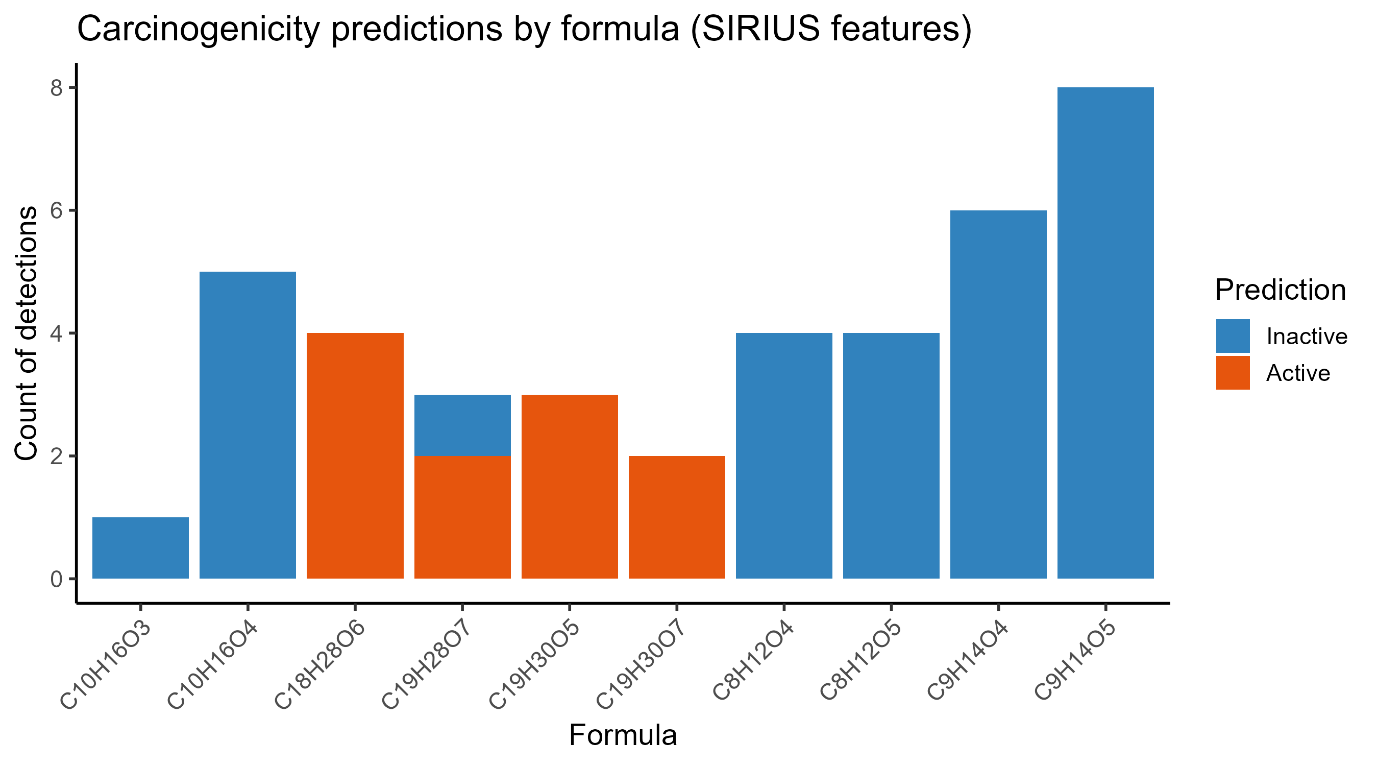
**Figure 2.5.4** SHAP beeswarm summary plot with SMILES annotations for the top 25 features, linking structural motifs to predicted carcinogenicity.

**3. Results: Deployment of the Model to Real Laboratory SOA Samples**  
  
This stage of the research was designed to explore if the built classifier could extend beyond idealised datasets and provide meaningful toxicological interpretations for real secondary organic aerosols (SOA). The deployment is essential since without it, the model could stay limited to idealised data, and there would be little application relevance for real environmental monitoring or chemical risk prioritisation.

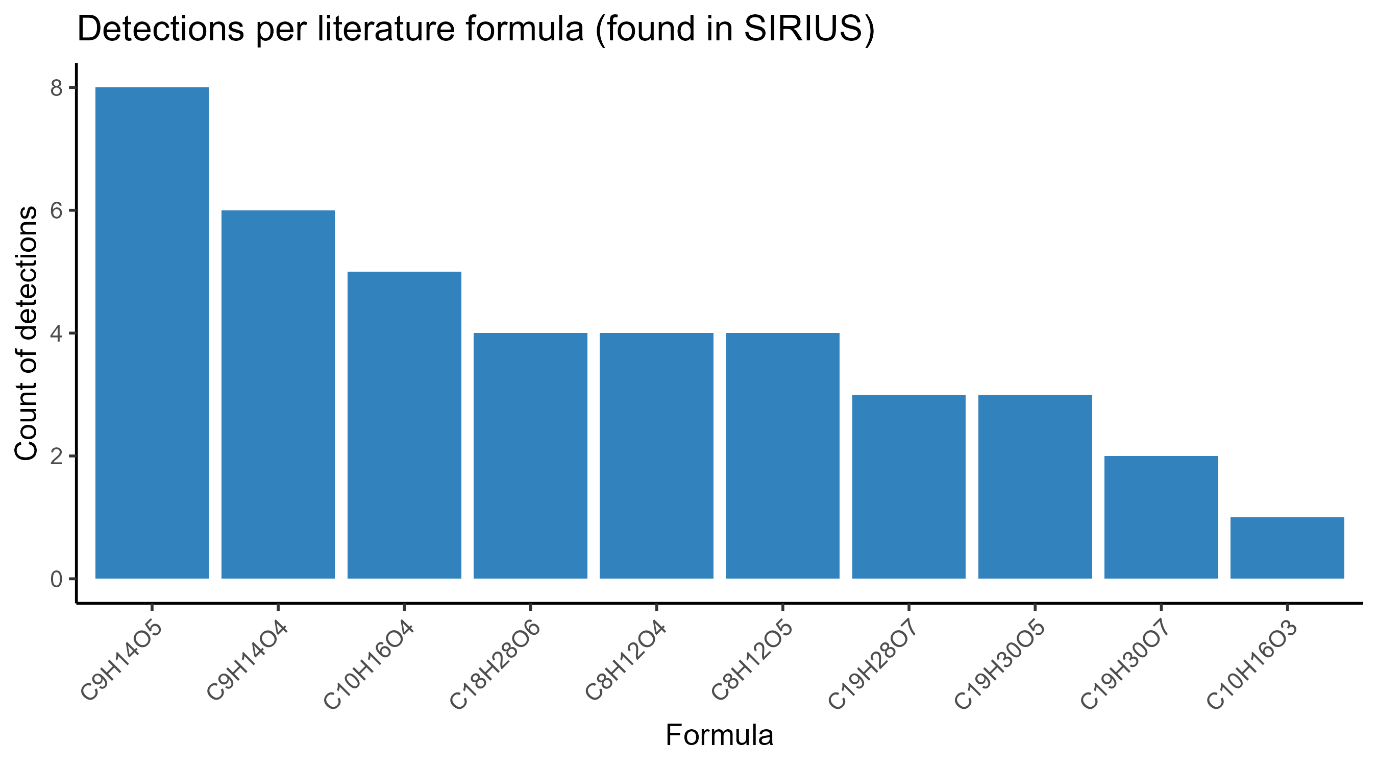
**3.1 Sample Generation and Analysis**  
A typical monoterpene precursor, limonene, was selected for its widespread use and high potential in forming SOA through ozonolysis, a key atmospheric oxidant (Bateman et al., 2009). Chamber tests were carried out under oxidative conditions using ozone, a typical means to mimic monoterpene atmospheric chemistry (Hallquist et al., 2009; Witkowski and Gierczak, 2017). Aerosols generated in experiments were retrieved on Teflon filters, solvent-extracted, and quantified by LC–MS/MS. This ensured that compounds identified were representative of the multi-generational limonene oxidation chemistry for realistic atmospheric conditions (Ehn et al., 2014; Kristensen et al., 2020).

**3.2 Inference of Molecular Formula and Fingerprint**

Raw data spectra were processed in SIRIUS v5.6.x, which assigns molecular formula through accurate mass, isotope pattern distribution, and fragmentation tree score (Böcker and Dührkop, 2016; Dührkop et al., 2019). Substructural fingerprints were subsequently predicted using the CSI:FingerID module, which calculates probabilistic descriptors from fragmentation patterns in a direct approach. Fingerprints were projected to the cheminformatics feature schema used in model training.  
  
Because limonene SOA is chemically complex, the same elemental composition would typically appear as more than one chromatographic peak of different retention times. Such an occurrence happens because of isomerism, whereby compounds sharing the same composition are of different polarity and connectivity that influences separation (Kopka et al., 2004).  
  
**3.3 Model Predictions on Real Sample Data**  
  
426 annotated formulas were discovered in the LC–MS/MS data. Each of them was assigned a probability of p53-pathway activation by the cross-validation-trained XGBoost model, and this was then converted to binary Active/Inactive calls utilising the ROC-optimised threshold (0.4805) learned under cross-validation (Figure 3.1).

  
**Figure 3.1 –** Predicted carcinogenicity (Active or Inactive) for literature-reported limonene SOA formulas detected in this study, based on SIRIUS-derived molecular fingerprints and the trained XGBoost classifier.

The majority of compounds were predicted as Inactive, with a minority small number labelled Active. Surprisingly, Active predictions were the majority among species of higher molecular weight and oxygen content in accordance with the hypothesis that electrophilic, multifunctional organics are more biologically reactive (Schwöbel et al., 2011). Smaller monomers with fewer oxygen atoms were predictably systematically classified as Inactive.  
  
**3.4 Literature Cross-Referencing**  
  
To provide perspective, annotated formulas documented here were compared with those already reported in chamber-generated limonene SOA (Witkowski and Gierczak, 2017). Level comparisons of the elemental formula facilitated direct comparison of model predictions with literature-reported species.



**Figure 3.2 –** Detection frequency of literature-reported limonene SOA formulas found in SIRIUS-processed LC–MS/MS data. Bars represent the number of detections per molecular formula, irrespective of predicted carcinogenicity.

As Figure 3.1 indicates, high-mass formulas such as C18H28O6, C19H28O7, and C19H30O5 were often tagged as Active, whereas low-mass molecules such as C8H12O4 and C10H16O4 were consistently Inactive. Figure 3.2 displays detection frequencies and indicates that prevalent species such as C9H14O5, C9H14O4, and C10H16O4 were predicted Inactive, while rarer, oxidised oligomers were often predicted Active. These trends indicate the manner in which structural elaboration, and not abundance, determines toxicological potential within SOA mixtures.  
  
**3.5 Importance and Limitations**  
This proof on actual samples demonstrates that a machine learning model, which is trained only on expertly handpicked toxicology data, can be used on unlabelled environmental mass-spectral data. The approach is the integration of structural inference (SIRIUS), substructure fingerprint mapping (CSI:FingerID), and probabilistic classification (XGBoost) and enables prioritisers of potentialally harmful compounds from hundreds of SOA features identified.  
  
The innovation is in the direct application of a carcinogenicity classifier to MS-inferred fingerprints from SOA produced in the chamber. To our knowledge, this is the first report on such an exercise. Notably, the reliability of this application is borne out by the preceding external validation (Section 2.5), where it was shown that the model is reliable in transitioning between curated structural data and spectral fingerprints.  
  
The results of our model should be interpreted cautiously, as predictive outputs rather than confirmed toxicological effects. Structural inferences and fingerprint estimations from non-targeted spectra introduce significant uncertainty particularly for isomeric or uncharacterised species. Machine learning systems lacking mechanisms for uncertainty quantification can produce overconfident predictions. These methods are thus best suited as triaging tools to flag potentially hazardous compounds for further experimental verification, not as substitutes for definitive hazard identification (Lazic and Williams, 2021; Sobus et al., 2018).  
  
By integrating controlled SOA formation with computational toxicology, the current study demonstrates a scalable path towards environmental monitoring and prioritization of potentially carcinogenic aerosols. It integrates atmospheric chemistry, analytical mass spectrometry, and predictive toxicology to advance application of New Approach Methodologies (NAMs) for chemical risk assessment (Escher et al., 2022; Thomas et al., 2019).

**4. Discussion**

The research was designed to determine whether machine learning would provide a mechanistically based and robust system for predicting the carcinogenic potential of ambient organic chemicals. The findings illustrate that an XGBoost classifier trained using p53 high-throughput assay data not only performs well for curated chemical structures, but generalizes well to spectral fingerprints and ultimately complex secondary organic aerosol (SOA) mixtures. In doing this, the research bridges an important divide between computationally derived toxicological screening and atmospheric measurement, showing that computational toxicology can be achieved under ambient environmental conditions.

Model performance with curated CompTox data highlights the importance of cheminformatics feature engineering. Structural fingerprints and monoisotopic mass alone were sufficient to capture patterns for p53 pathway activation, with balanced accuracies of >90% under ROC-optimised thresholds. Impressively, this predictive power was not lost when translated to the more challenging space of spectral fingerprints of MS/MS data. Although accuracy fell to ~72% in external validation, this is consistent with the loss of information that is expected when translating from intact molecular graphs to fragmentation-based descriptors (Dührkop et al., 2019; Peets et al., 2022). Rather than the signature of model failure, this dampening makes the method predicatively valuable even in domain shift the need for environmental deployment where spectra overwhelm chemical detection.

Example with limonene SOA showed an even more significant finding: model predictions weren't random but followed chemical intuition. Bigger, oxygenated molecules were disproportionately active and smaller, less functionalised monomers were always inactive. This is consistent with mechanistic theory that more multifunctional oxidised organics are biologically more active, echoing also earlier toxicological evidence (Schwöbel et al., 2011). That such domain-specific trends in atmospheric chemistry can be mined by a machine learning system trained solely on general toxicology data speaks to its explanatory and real-world relevance.

Caveats moderate these developments at the same time. The CompTox dataset remained imbalanced with relatively few active compounds, and this is certain to have biased the classifier towards the majority class despite weighting schemes. In addition, XGBoost by itself was extensively piloted; although its interpretability and usefulness render it a sound choice, comparison trials with alternative models such as Random Forests, Support Vector Machines, or novel graph neural networks would strengthen claims of improved generalisability (Limbu and Dakshanamurthy, 2022; Mayr et al., 2016). Validation was also constrained by reference data available: the MassBank collection, though of high quality, represents but a subset of atmospheric chemical diversity. Likewise, the SOA application is predictive, rather than confirmatory, since ground truth toxicology does not exist for most formulae detected. Inference of fingerprint from spectra represents a source of uncertainty, particularly for isomers not distinguishable by mass.

Despite these limitations, the study has profound implications. This also serves the purpose of being able to use it as a triage tool, pushing environmental chemicals to regulatory priority and into the lab. With less than 5% of air chemicals having toxicological profiles (Thomas et al., 2019), being able to scale predictions ethically and cost-effectively is of tremendous value. By reducing the need for animal tests and complying with New Approach Methodologies (OECD, 2016), the system provides scientific progress as well as societal benefit. Moreover, the union of predictive toxicology and atmospheric chemistry highlights the value of cross-disciplinary interactions, advancing air quality science, public health, and regulatory decision-making in tandem.

Future research should broaden methodological and chemical scope. Incorporating several mechanistic assays in addition to p53 would allow multi-endpoint models more representatively capturing the complexity of carcinogenic processes. Validation to more SOA precursors and to ambient air samples would challenge robustness even more. More importantly, the inclusion of interpretable methods such as SHAP would be able to reveal what structural motifs are driving predictions, tracing computational outputs back to mechanistic toxicology (Lundberg and Lee, 2017; Walter, 2022). By combining methodological development with area expansion, future work can render computational prediction of carcinogenicity a cornerstone of environmental health science.

Briefly, this study illustrates the potential for using machine learning in the prediction of environmental carcinogens in real chemical mixtures. Its utility is less in the performance metrics and more in providing a path forward: one that uses computational power to enhance experimental effort, one where air pollutants are screened more efficiently and ethically, and one where regulators can make better-informed decisions in the context of chemical uncertainty.

**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 and powerfully to the study of atmospheric carcinogens, both in a scientific and in a potential public health protection sense.

**6. References**

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

Dix, D.J., Houck, K.A., Martin, M.T., Richard, A.M., Setzer, R.W., Kavlock, R.J., 2007. The ToxCast Program for Prioritizing Toxicity Testing of Environmental Chemicals. Toxicol. Sci. 95, 5–12. https://doi.org/10.1093/toxsci/kfl103

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

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

Ehn, M., Thornton, J.A., Kleist, E., Sipilä, M., Junninen, H., Pullinen, I., Springer, M., Rubach, F., Tillmann, R., Lee, B., Lopez-Hilfiker, F., Andres, S., Acir, I.-H., Rissanen, M., Jokinen, T., Schobesberger, S., Kangasluoma, J., Kontkanen, J., Nieminen, T., Kurtén, T., Nielsen, L.B., Jørgensen, S., Kjaergaard, H.G., Canagaratna, M., Maso, M.D., Berndt, T., Petäjä, T., Wahner, A., Kerminen, V.-M., Kulmala, M., Worsnop, D.R., Wildt, J., Mentel, T.F., 2014. A large source of low-volatility secondary organic aerosol. Nature 506, 476–479. https://doi.org/10.1038/nature13032

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

Fernandez-Delgado, M., Cernadas, E., Barro, S., Amorim, D., n.d. Do we Need Hundreds of Classiﬁers to Solve Real World Classiﬁcation Problems?

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

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

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

Hutter, F., Kotthoff, L., Vanschoren, J. (Eds.), 2019. Automated Machine Learning: Methods, Systems, Challenges, The Springer Series on Challenges in Machine Learning. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-05318-5

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

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

Kopka, J., Fernie, A., Weckwerth, W., Gibon, Y., Stitt, M., 2004. Metabolite profiling in plant biology: platforms and destinations. Genome Biol. 5, 109. https://doi.org/10.1186/gb-2004-5-6-109

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

Kristensen, K., Jensen, L.N., Quéléver, L.L.J., Christiansen, S., Rosati, B., Elm, J., Teiwes, R., Pedersen, H.B., Glasius, M., Ehn, M., Bilde, M., 2020. The Aarhus Chamber Campaign on Highly Oxygenated Organic Molecules and Aerosols (ACCHA): particle formation, organic acids, and dimer esters from *α*-pinene ozonolysis at different temperatures. Atmospheric Chem. Phys. 20, 12549–12567. https://doi.org/10.5194/acp-20-12549-2020

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., n.d. p53, the Cellular Gatekeeper Review for Growth and Division.

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

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, 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

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

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

WHO, 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).

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).

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

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**