**From Structures to Risk: Predicting p53-Mediated Carcinogenicity in Atmosphere Chemical Using Machine Learning**

**1. Introduction**

**1.1 Air Pollution and Hidden Cancer Risks**

Air pollution is among the leading environmental threats to public health, contributing to an estimated seven million premature deaths annually (WHO, 2023). Fine particulate matter, especially particles less than 2.5 microns in diameter (PM2.5), poses a unique risk due to its ability to penetrate deep into the lungs and enter systemic circulation (Brook et al., 2010). Once distributed through the body, these particles can interfere with biological processes and increase disease susceptibility.

PM2.5 is hazardous not just because of its size but due to its chemical complexity. It can carry toxic compounds such as polycyclic aromatic hydrocarbons (PAHs), volatile organic compounds (VOCs), nitroaromatics, transition metals like lead and cadmium, and reactive oxygen species (ROS), all of which are implicated in carcinogenesis (Shrivastava et al., 2017; Kelly and Fussell, 2012). These pollutants primarily originate from traffic emissions, fossil fuel combustion, industrial discharge, and agricultural biomass burning (Hallquist et al., 2009; Seinfeld and Pandis, 2016).

The combination of particle size and chemical toxicity renders PM2.5 a significant contributor to chronic illnesses including respiratory and cardiovascular diseases, and notably cancer (Shiraiwa et al., 2017; Zhang et al., 2018). Yet, despite the scale of this threat, less than 5% of airborne chemicals have undergone rigorous long-term toxicity testing (EPA, 2020; Thomas et al., 2019). This lack of toxicological characterisation severely limits public health protection and regulatory response.

**1.2 Understanding Cancer and the Role of the p53 Pathway**

Carcinogenesis involves the accumulation of mutations that disrupt cellular regulatory mechanisms, enabling uncontrolled growth, immune evasion, and survival despite DNA damage (Hanahan and Weinberg, 2011). A central defender against such transformation is the p53 protein, a tumour suppressor that orchestrates responses to DNA damage, including cell cycle arrest, apoptosis, and repair mechanisms.

p53 is normally inactive but is rapidly activated in response to cellular stress. It can then halt cell division, initiate DNA repair, or trigger programmed cell death, thereby preventing the propagation of damaged cells (Lane, 1992; Levine, 1997). Mutations in TP53, the gene encoding p53, or its regulatory pathways are observed in more than half of all human cancers (Olivier et al., 2010; Kastenhuber and Lowe, 2017).

Many environmental carcinogens, including those found in air pollution, disrupt p53-mediated pathways, either by inducing DNA damage or by inhibiting p53 function (Hussain and Harris, 2007; Zeron-Medina et al., 2013). This makes the p53 pathway an effective target for identifying potential chemical carcinogens.

High-throughput assays like ATG\_p53\_CIS, developed by the U.S. Environmental Protection Agency and the National Center for Advancing Translational Sciences, allow for rapid screening of chemicals. Using a fluorescent reporter system in engineered human cells, the assay provides a binary readout “active” or “inactive” reflecting whether a compound perturbs the p53 pathway (Judson et al., 2010; Richard et al., 2016). This approach enables mechanistic toxicity screening at scale.

**1.3 Traditional and Computational Approaches to Chemical Risk Assessment**

Traditional toxicology relies heavily on animal testing, particularly rodent bioassays that span months or years and evaluate cancer development after chemical exposure (Zeiger, 2019; Krewski et al., 2010). While these studies provide critical data, they are time-consuming, expensive, and raise ethical concerns due to extensive animal use. Additionally, interspecies differences limit their applicability to human health (Basketter et al., 2012).

Another limitation is the narrow scope: animal studies typically test single chemicals in isolation, whereas humans are exposed to complex mixtures. The CompTox Chemicals Dashboard contains over 875,000 entries, yet the majority lack detailed toxicological profiles (Williams et al., 2017). As a result, new approach methodologies (NAMs) are gaining traction for their ability to offer faster, mechanistically relevant, and animal-free alternatives (OECD, 2016).

These approaches include high-throughput in vitro assays, organ-on-chip models, and computational tools that predict toxicity based on molecular structure and bioactivity. Programmes such as ToxCast and Tox21 have applied these methods to thousands of compounds across multiple biological pathways, including carcinogenesis (Dix et al., 2007; Kavlock et al., 2009).

Among computational methods, early models used logistic regression, random forests, and support vector machines (SVMs). Each has strengths logistic regression is interpretable; SVMs handle complex boundaries; and random forests reduce variance but all struggle with scaling and transparency in large chemical datasets.

Recent innovations include deep learning and ensemble methods such as XGBoost, which offer improved performance, especially in high-dimensional settings. Graph neural networks (GNNs) represent a new frontier by learning directly from molecular graphs, although their use in toxicology is still emerging. Interpretability tools such as SHAP (Shapley Additive Explanations) have also improved transparency by assigning importance scores to individual features (Lundberg and Lee, 2017).

Together, these tools expand the toxicologist’s toolkit, enabling faster and more scalable assessments of chemical risk.

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

Machine learning (ML) now plays a pivotal role in computational toxicology by enabling predictions based on chemical features such as structure, physicochemical properties, and assay outputs. Known as in silico toxicology, these models offer a scalable, cost-effective alternative to traditional experiments.

XGBoost, a popular gradient boosting algorithm, is particularly well-suited for such tasks due to its speed, handling of missing data, and strong performance in imbalanced datasets (Chen and Guestrin, 2016). When combined with SHAP, XGBoost models become more interpretable, allowing researchers to trace how specific chemical features influence toxicity predictions.

Several studies illustrate the effectiveness of ML in toxicology. Cornell et al. (2022) applied XGBoost to predict carcinogenicity in aerosols using LC-MS data. Peets et al. (2022) developed MS2Tox, which estimates toxicity based on MS2 fragmentation. Dührkop et al. (2019) introduced CSI:FingerID and SIRIUS, which infer structure and activity from MS/MS spectra, supporting predictions even when full molecular identities are unknown. More recently, Cai et al. (2024) identified novel carcinogens in ambient air using similar ML frameworks, while Arturi and Hollender (2023) used Random Forests to rank chemicals for regulatory prioritisation.

These studies show how ML can fill gaps left by conventional toxicology, offering data-driven predictions that support regulatory screening and hypothesis generation.

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

Predictive models are only useful if they work in real-world settings. In environmental toxicology, this means validating model predictions using compounds identified in the atmosphere often through non-targeted analysis with techniques such as LC-MS/MS (Kind and Fiehn, 2010).

LC-MS/MS allows researchers to separate, detect, and characterise thousands of compounds in air samples. Secondary organic aerosols (SOAs), which are major components of PM2.5, can be profiled using this approach (Shrivastava et al., 2017; Witkowski and Gierczak, 2017). However, many compounds detected are uncharacterised and not found in regulatory databases.

To interpret these spectra, tools like SIRIUS and CSI:FingerID generate molecular formulas and predict substructural fingerprints based on fragmentation patterns (Dührkop et al., 2015; 2019). These predicted fingerprints can then be fed into ML models trained on known compounds to predict the toxicity of unknowns.

Cornell et al. (2022) and Peets et al. (2022) demonstrated this approach by using XGBoost and MS2Tox to infer toxicity from spectral fingerprints. This method extends toxicity predictions to chemicals not yet catalogued, supporting real-time chemical surveillance and prioritisation.

In this study, a similar strategy is employed: spectral data from the LCSB MassBank is processed using SIRIUS, and predicted fingerprints are used as input to the trained XGBoost model to infer potential p53 pathway activation. This allows validation against real environmental exposures and not just curated datasets.

**1.6 Why This Research Matters: Safer Air Through Smarter Models**

Air pollution contains thousands of chemicals, but most have never been tested to see if they can cause cancer (EPA, 2020; Thomas et al., 2019). Traditional testing on animals takes years, is expensive, and raises ethical concerns (Zeiger, 2019; Basketter et al., 2012). Newer approaches, such as high-throughput screening (HTS) and computer-based toxicology, can test chemicals faster, on a larger scale, and without relying heavily on animal experiments (OECD, 2016; Jaworska and Hoffmann, 2010).

In this study, a machine learning model called XGBoost is used to predict whether a chemical might interfere with a cancer-related biological pathway, based on its molecular “fingerprints” and high-throughput test results. Molecular fingerprints are numerical codes that describe a molecule’s structure so it can be analysed by a computer (Guha, 2007). High-throughput assays are automated lab tests that quickly measure how many chemicals affect a certain biological process (Kavlock et al., 2009).

To make sure the model works in real-world situations, it is also tested on data from environmental samples analysed using tandem mass spectrometry (MS/MS). In these cases, chemical structures are predicted from spectral patterns using tools like CSI:FingerID (Dührkop et al., 2019; Peets et al., 2022), similar to how pollutants are detected in air monitoring studies.

This approach supports important safety and ethics frameworks, including Next Generation Risk Assessment (NGRA) and the 3Rs principle — Replacement (avoiding animal testing where possible), Reduction (using fewer animals), and Refinement (making tests less harmful) (Russell and Burch, 1959; Thomas et al., 2019). The method is designed to be reproducible and adaptable, helping scientists and regulators identify high-risk air pollutants more quickly, especially in areas with high exposure from traffic or industry (Landrigan et al., 2018; Arturi and Hollender, 2023).

**2. Data, Tools and Machine Learning Methodology**

This section provides a comprehensive explanation of the data sources, cheminformatics feature engineering, machine learning strategy, and external validation procedures applied to construct a robust and interpretable predictive model for chemical carcinogenicity. The overall goal of this study is to predict the likelihood that an atmospheric organic compound activates the p53 tumour suppressor pathway, a hallmark of genotoxic and carcinogenic processes. The pipeline integrates publicly available high-throughput toxicological assays with cheminformatics-based structural descriptors, advanced machine learning techniques, and external spectral fingerprint validation. All work was conducted using R version 4.3.1 within an open-source, reproducible computational environment.

**2.1 Data Acquisition and Biological Justification**

The primary data source for this study was the EPA CompTox Chemicals Dashboard, which aggregates high-throughput screening (HTS) data from the ToxCast and Tox21 programs. Specifically, the ATG\_P53\_CIS assay was selected, which monitors transcriptional activity related to the tumour suppressor protein p53. This protein plays a central role in DNA repair, cell cycle arrest, and apoptosis and is widely regarded as the “guardian of the genome” (Levine, 1997). Mutations or disruptions in p53 are implicated in over 50% of all human cancers (Vousden and Prives, 2009), making it a critical biomarker in toxicological screening.

From the CompTox dataset, an initial pool of 4039 compounds was extracted, comprising both active and inactive labels based on their ability to induce a signal in the ATG\_P53\_CIS assay. Only organic compounds with valid SMILES strings, CAS numbers, and monoisotopic mass information were retained. Further filtering was conducted to exclude duplicates, mixtures, and compounds lacking defined molecular formulae. To ensure relevance to atmospheric chemistry, an additional filter was applied to retain only compounds containing C, H, N, O, and S atoms, reflecting common constituents of airborne organics (Shrivastava et al., 2017).

After removing entries with missing data and applying rigorous structural and assay-based criteria, a final cleaned dataset of 718 compounds was obtained, including 127 actives and 591 inactives. These compounds formed the basis of all downstream cheminformatics processing and model development.

**2.2 Cheminformatics Feature Generation**

To convert chemical structures into machine-readable inputs, a wide array of molecular fingerprints was computed using the rcdk package in R (Guha, 2007). Fingerprints are binary or numeric representations that capture the presence or absence of substructural motifs, topological features, and physicochemical properties.

Several types of fingerprints were used in this study to ensure a rich and diverse representation of structural information. These included MACCS keys (166-bit substructure patterns), PubChem fingerprints (881-bit features used by the PubChem database), and Klekota-Roth fingerprints (4860-bit patterns optimised for bioactivity prediction). Substructure-based fingerprints such as the CDK standard (1024-bit) and EState fingerprints were also used to capture electronic and topological properties.

In addition to these canonical descriptors, custom SMARTS-based fingerprints were generated using specific reactive substructures relevant to carcinogenicity. Furthermore, ring system counts, atom-type distributions, and monoisotopic mass were appended to the final feature matrix. This comprehensive approach ensured that the model received both general and specific chemical information potentially related to carcinogenic activity.

All fingerprint vectors were normalised and aligned to retain only common features across compounds. Features with near-zero variance were removed to reduce noise, and highly correlated features (correlation > 0.9) were eliminated to prevent multicollinearity. These preprocessing steps are essential in cheminformatics to ensure model stability and generalisability (Svetnik et al., 2003).

**2.3 Data Preprocessing and Feature Selection**

Data preprocessing was conducted using the caret and dplyr packages. After computing all cheminformatics features, missing values were removed and a complete cases dataset was retained. Monoisotopic mass was scaled using z-score standardisation to bring all variables onto a similar range.

Near-zero variance features were identified using the nearZeroVar function in caret. These features contribute minimal variation and can reduce model efficiency, especially in tree-based algorithms. Correlated features were pruned using Pearson correlation matrices to reduce redundancy and avoid overfitting.

The binary outcome variable, HIT.CALL, was converted into a numeric vector with 1 representing actives and 0 representing inactives. To correct for class imbalance, the ratio of inactive to active compounds was used to compute a scale\_pos\_weight parameter for the XGBoost classifier. Class imbalance is a common problem in toxicology datasets, and weighting approaches help to avoid bias toward the majority class (Chawla et al., 2002).

Train/test splits were performed using an 80:20 stratified partition to maintain the proportion of actives in both sets. These subsets were used during model tuning and cross-validation to evaluate generalisability. Data matrices were converted into sparse matrices using xgb.DMatrix for computational efficiency.

**2.4 Machine Learning Model Development Using XGBoost**

To predict carcinogenic activity from structural fingerprints, the gradient boosting algorithm XGBoost (Extreme Gradient Boosting) was selected as the core modelling framework. XGBoost is an ensemble learning technique that builds a sequence of decision trees, where each subsequent tree corrects the prediction errors of the previous ones (Chen and Guestrin, 2016). This method is particularly well-suited for high-dimensional, imbalanced datasets like those encountered in cheminformatics, due to its scalability, regularisation capabilities, and performance across classification tasks (Sheridan, 2013; Mayr et al., 2016).

XGBoost optimises a differentiable loss function through gradient descent, using second-order approximations to speed up convergence. It incorporates L1 and L2 regularisation, column and row subsampling, and learning rate shrinkage to reduce overfitting. Parameters such as gamma (minimum loss reduction required to make a further partition) and eta (learning rate) help balance model complexity and generalisation. The colsample\_bytree and subsample parameters reduce reliance on specific features or samples, enhancing model robustness (Probst et al., 2019).

Model development initially began with 10-fold cross-validation using the caret package in R, applying a grid of hyperparameters. However, this setup proved time-consuming and did not yield improved accuracy. To address this, the workflow was revised to use 5-fold cross-validation, a strategy that is often sufficient in toxicology modelling and reduces runtime significantly (Boulesteix et al., 2017). Rather than grid-searching all parameters together, models were trained independently with different values for nrounds (number of boosting iterations) to enable comparative evaluation. This modular approach increased interpretability and computational efficiency (Hutter et al., 2019).

The best-performing model on training data used nrounds = 100, eta = 0.01, max\_depth = 6, gamma = 5, colsample\_bytree = 0.6, and subsample = 0.6. After selecting this configuration, the final model was retrained on the entire dataset (718 compounds) to ensure that the learning algorithm benefited from all available structural and toxicity data. This practice is recommended in machine learning literature for small to medium-sized datasets where generalisability is more important than hold-out validation (Kuhn and Johnson, 2013).

Instead of relying on the default classification threshold (0.5), the study used Youden’s J statistic to identify the optimal threshold from the ROC curve. This method is well-accepted in medical diagnostics and binary classification tasks because it balances sensitivity and specificity (Youden, 1950; Powers, 2011).

Two thresholds were examined to demonstrate the impact on classification performance. When the decision threshold was set to 0.47, the model prioritised sensitivity and identified 118 true positives out of 127, yielding a sensitivity of 92.91% and a balanced accuracy of 82.58%. However, this came with 164 false positives and a lower precision of 41.84%, illustrating a classic trade-off between recall and specificity. This configuration is particularly useful for applications where it is more dangerous to miss a positive case, such as cancer screening or toxicological hazard identification (Fernandez-Delgado et al., 2014; Chawla et al., 2002).

In contrast, using a stricter threshold of 0.65 reduced false positives to 40, leading to a much higher specificity of 93.23% and precision of 65.22%. However, this came with a drop in sensitivity to 59.06%, detecting only 75 of the 127 active compounds. The overall accuracy increased to 87.19%, demonstrating that this threshold configuration is better suited for use cases where false positives are costly, such as regulatory decision-making or prioritised chemical screening (Hastie et al., 2009).

Following threshold optimisation, the final XGBoost model was trained on the full dataset using the selected hyperparameters. The ROC-derived optimal threshold for this retrained model was 0.4805. At this setting, the model achieved an accuracy of 90.25%, sensitivity of 93.70%, specificity of 89.51%, and balanced accuracy of 91.61%. The Kappa statistic was 0.7131, indicating strong agreement between predicted and actual labels beyond chance (Cohen, 1960; Brodersen et al., 2010). These metrics confirm that the model was able to generalise well across both active and inactive classes.

All performance evaluations were carried out using the pROC and caret packages in R. Confusion matrices were generated, and results were visualised with ROC curves and performance metrics. Model reproducibility was ensured by fixing random seeds and saving trained model objects using saveRDS. This approach aligns with best practices in computational modelling and reproducibility (Sandve et al., 2013; Wilkinson et al., 2016).

Every modelling decision, including parameter tuning, evaluation metric selection, and threshold choice, was guided by established practices in cheminformatics and toxicological data science. The deliberate structuring of training stages, use of ROC-based optimisation, and retraining on combined data aimed to create a model that is not only performant but also adaptable to different real-world scenarios in chemical safety assessment.

**2.5 External Validation Using Mass Spectral Data**

To enhance the real-world relevance and generalisability of the model, external validation was conducted using mass spectrometry-derived compound data curated from the Luxembourg Centre for Systems Biomedicine (LCSB) MassBank. This dataset comprises experimentally acquired MS/MS spectra for thousands of environmental and biological compounds. The aim of this validation was to assess whether the model, trained on structural fingerprints, could effectively predict the carcinogenic potential of previously unseen compounds whose structures were confirmed through high-resolution tandem mass spectrometry.

The LCSB dataset initially contained over 5,500 compounds. To ensure comparability with the training data, a series of strict filtering criteria were applied. First, only compounds with complete metadata, valid CAS Registry Numbers (CASRN), and experimentally confirmed structures were retained. These were then cross-referenced with the curated list of compounds from the U.S. EPA CompTox Dashboard, matching on CASRN to ensure label compatibility with the p53 activity data. This resulted in a final validation set of 319 compounds with high-confidence spectral and structural alignment.

Fingerprint prediction for these compounds was achieved using the CSI:FingerID module within SIRIUS v5.6.3. This tool predicts molecular fingerprints directly from MS/MS spectra using kernel support vector machines trained on large chemical structure databases (Dührkop et al., 2015; Dührkop et al., 2019). These predictions mimic the types of structural fingerprints used in the training data but are inferred from spectral features rather than 2D molecular structures. This indirect method of fingerprint generation allows for an assessment of model robustness when fingerprints are obtained from experimental data rather than cheminformatics calculations.

Once the predicted fingerprints were generated, they were processed to match the format and feature space of the original training dataset. The XGBoost model trained on structural fingerprints was then applied to these spectra-inferred fingerprints, generating predicted probabilities of p53 activity for each compound. Class assignments were made using the previously determined ROC-optimised threshold.

Validation performance was assessed by comparing the model’s predictions to the known p53 activity labels from the CompTox dataset. A confusion matrix was constructed, and performance metrics such as accuracy, sensitivity, specificity, and precision were calculated. This evaluation provided insight into the model’s ability to generalise across different types of input data, confirming its potential utility for chemical screening even when traditional structural representations are unavailable or incomplete.

The use of spectral validation represents a significant step toward regulatory applicability and environmental deployment. Many atmospheric or industrial chemicals exist only in complex mixtures and are characterised through high-resolution mass spectrometry rather than full structure elucidation. The ability of the model to accurately classify these compounds from spectral data alone supports its use in real-world chemical monitoring and prioritisation tasks (Peets et al., 2022; Arturi and Hollender, 2023).

Moreover, this approach aligns with modern principles of ethical and sustainable toxicology, reducing reliance on animal testing by integrating computational predictions with high-throughput analytical chemistry data. By validating the model against independent, experimentally derived fingerprints, this study demonstrates the feasibility of using spectral information for toxicological classification, thereby extending the practical value of machine learning in chemical risk assessment.

All data preprocessing, fingerprint alignment, and prediction steps were performed using reproducible R scripts. The results from this external validation phase further confirm the utility and adaptability of the model across diverse data modalities and reinforce its robustness for downstream applications in environmental health research.

**2.6 Summary of Methodological Contributions and Limitations**

The methodology adopted in this study integrates modern cheminformatics with interpretable machine learning to construct a robust framework for predicting chemical carcinogenicity. By leveraging high-quality data from the U.S. EPA CompTox Dashboard and advanced fingerprinting tools in R, the workflow ensured both reproducibility and chemical relevance. The choice of XGBoost as the classification algorithm was justified by its strong performance on high-dimensional, imbalanced datasets, typical in chemical hazard prediction (Chen and Guestrin, 2016; Sheridan, 2013).

One of the study’s key contributions lies in its flexible modelling pipeline that allows for both structure-based and spectrum-based prediction of chemical activity. This is particularly valuable in environmental monitoring, where many pollutants are detected through mass spectrometry rather than fully elucidated chemical structures. The use of CSI:FingerID for fingerprint inference further extends the model’s applicability to experimentally profiled but structurally unannotated compounds (Dührkop et al., 2019).

Threshold optimisation using ROC analysis enabled the construction of models tuned for different decision contexts. For example, a lower threshold emphasised sensitivity—ideal for screening applications where missing a carcinogen is unacceptable—while a higher threshold improved specificity, appropriate for regulatory or prioritisation scenarios where false positives are costly. This dual-mode interpretability is a significant methodological asset.

However, the study is not without limitations. The dataset, although curated, still contains a degree of label noise due to the inherent complexity and uncertainty of bioassay outcomes. Moreover, while external validation was performed, the predicted fingerprints from CSI:FingerID may not fully capture the intricacies of 2D structural fingerprints, potentially introducing a source of variability. Future work should explore ensemble predictions that combine multiple fingerprint types and assess broader generalisability across additional chemical classes and assays.

Despite these limitations, the integrated and reproducible methodology developed here represents a scalable and ethically conscious approach to predictive toxicology. It aligns with current trends in regulatory science and computational chemistry, offering a versatile tool for prioritising hazardous chemicals based on their potential to activate a key cancer-related biological pathway.

Results