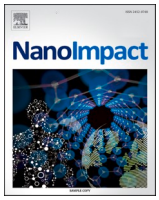
NanoImpact 38 (2025) 100562

Contents lists available at ScienceDirect

NanoImpact

journal homepage: www.elsevier.com/locate/nanoimpact



Research Paper

From literature to predictive modeling: Insights and machine learning applications from in vitro comet assays related to the genotoxicity of titanium dioxide nanomaterials

Irini Furxhi [a,1,*], Mahsa Mirzaei [b,1], Anna Costa [a,*], Rossella Bengalli [c]

[a] CNR-ISSMC Istituto di Scienza e Tecnologia dei Materiali Ceramici, Via Granarolo, 64, 48018 Faenza, RA, Italy
[b] School of Biomolecular and Biomedical Science, UCD Conway Institute, University College Dublin, Dublin, Ireland
[c] Department of Earth and Environmental Sciences, University of Milano-Bicocca, Piazza della Scienza 1, Milano 20126, Italy

ARTICLE INFO

Editor: Phil Demokritou

Keywords:
Titanium dioxide (TiO_2)
Genotoxicity
Nanomaterials
Nanoforms
Machine learning
NAMs

ABSTRACT

The genotoxicity of titanium dioxide nanomaterials (TiO_2 NMs) remains a debated topic in the scientific community. In this study, we applied the read-across concept based on machine learning (ML) algorithms to predict the genotoxic potential of TiO_2 NMs. Key objectives included: (i) compiling a systematic dataset capturing DNA damage percentage from in vitro comet assays, (ii) creating a homogenized dataset integrating physicochemical properties, exposure conditions, and experimental details, (iii) training ML models for prediction, (iv) evaluating model performance, and (v) identifying the features that contribute the most to predictive accuracy. The dataset was divided into three parts: the Entire dataset (all features), the Physicochemical dataset, and the Experimental design dataset. Extra Trees Regressor and XGB Regressor demonstrated high predictive accuracy, achieving R^2 values of 0.906 and 0.788 for the P-chem and Experimental dataset, respectively. Exposure concentration, cold lysis conditions, and electrophoresis parameters emerged as key predictors of DNA damage, alongside contributions from NM properties. These findings highlight the intricate interplay between NM properties and experimental conditions in genotoxicity assessments. By providing a FAIR dataset, this study facilitates future research, allowing for the integration of additional variables and quality criteria to enhance the modeling approach. This work reinforces the value of nano-informatics in nanosafety and serves as a footing for advancing data-driven hazard assessment methodologies, positioning ML-enabled read-across strategies as a valuable tool for regulatory nanosafety framework.

1. Introduction

Engineered nanomaterials (NMs) or nanoforms are characterized by particles in unbound, aggregated or agglomerated states with ≥50 % of particles having at least one external dimension within the 1–100 nm range (Commission, E, 2019). These materials exhibit unique physicochemical (pchem) properties distinguishing them from their bulk counterparts. Titanium Dioxide (TiO_2) is among the most widely produced NM, valued for its optical, chemical and photocatalytic properties, making it integral in coatings, sensors and electronic devices. It also serves as a prominent white pigment in cosmetics, paints, plastics and ceramics (Gázquez et al., 2021; Gázquez et al., 2014; Ziental et al., 2020). Its widespread application in consumer products, including food, personal care items (e.g., sunscreens, toothpaste, shampoos and deodorants) (Weir et al., 2012), has raised concerns about their safety. Notably, the potential genotoxicity of TiO_2-NMs has been subject to extensive safety evaluations (Carriere et al., 2020a). The European Food Safety Authority (EFSA) has deemed TiO_2, specifically grade E171, "no longer safe" for use in food due to evidence indicating its potential to induce DNA strand breaks and chromosomal damage, though not gene mutations; As a result, genotoxicity concerns could not be excluded (Additives et al., 2021). While its removal from food products has been implemented, a similar ban in pharmaceuticals could impact the industry, regulatory agencies, and patients. An estimated 91,000 approved drugs in Europe rely on TiO_2 as an excipient (Abend et al., 2024); however, it is important to note that TiO_2 exists in different forms, and

* Corresponding authors.
E-mail addresses: irini.furxhi@issmc.cnr.it (I. Furxhi), Mahsa.Mirzaei1@ucd.ie (M. Mirzaei), anna.costa@issmc.cnr.it (A. Costa), rossella.bengalli@unimib.it (R. Bengalli).
[1] These Authors contributed equally to this work.

https://doi.org/10.1016/j.impact.2025.100562
Received 21 December 2024; Received in revised form 20 March 2025; Accepted 21 April 2025
Available online 22 April 2025

not all of these pharmaceutical products would necessarily be affected by potential regulatory restrictions. Reformulation timelines are projected at 3–5 years per product, with regulatory approval taking an additional 3 months to 1 year. Considering the scale of products involved, industry estimates a transition period of 7–12 years and associated high costs reaching billions[2] (Teasdale and Hughes, 2023). To address these challenges, the Titanium Dioxide Manufacturers Association (TDMA) convened an independent expert panel to review existing data and identify gaps for further studies, aiming to provide the European Commission with evidence supporting the continued use of E171 in medicines. Totally separate from the food and pharmaceutical use of TiO_2, in occupational settings, TiO_2-NMs in powder form have been classified by the International Agency for Research on Cancer (IARC) as a Group 2B carcinogen (possibly carcinogenic to humans) via inhalation. However, a 2022 ruling by the European Court of Justice annulled this classification, citing errors in the reliability and acceptability of the supporting study and failure to meet criteria requiring intrinsic carcinogenic properties.[3] If upheld on appeal, this decision may allow manufacturers to remove warnings from product labels.

Despite differences in pchem properties and regulatory restrictions between TiO_2 and TiO_2-NMs, the particle size distribution in certain products may result in a fraction of NMs. A recent study (Liang et al., 2024) investigated the genotoxicity of a commercial food additive containing TiO_2 with an average particle size of 135 ± 41 nm (range: 60–230 nm), of which 30 % were NMs. Using a battery of standard in vivo genotoxicity tests, researchers administered intragastric doses of 250, 500, and 1000 mg/kg body weight over 15 days. The results indicated no increase in micronuclei or chromosomal aberrations in mouse bone marrow and no DNA strand breakage in rat liver cells. These findings suggest that, under the tested conditions, TiO_2 does not exhibit genotoxic potential, despite containing a nanoscale fraction. While the study emphasizes in vivo assessments, the extensive body of in vitro research also provides valuable insights. Although in vitro models cannot fully replicate in vivo conditions, they remain critical for understanding the potential genotoxic effects of TiO_2-NMs. Numerous in vitro studies have utilized assays to assess the DNA-damaging potential of TiO_2-NMs. These include tests such as the comet assay and its modified versions (e.g., Endo-III, hOGG1, and FPG-modified assays), as well as chromosome mutation tests like the flow-cytometry micronucleus (FCMN) and cytokinesis-block micronucleus (CBMN) assays. A gene mutation test (preferably in mammalian cells, such as the MLA (OECD TG 490; (OECD, 2016)) or the HPRT (OECD TG 476, (OECD, 2016a)) combined with a test for chromosome mutations (such as the CBMN (OECD TG 487, (OECD, 2023)) or the CA test (OECD TG 473, (OECD, 2016b)) may be recommended as members of a battery of in vitro NM genotoxicity tests. Nowadays the genotoxicity of NMs is still a matter of discussion: according to a recent expert consensus (Doak et al., 2023a), the standard genotoxicity tests originally designed for chemicals could have some limitations when applied *to* in vitro testing of NMs. In those tests are also included several standard OECD TG for genotoxicity, such as TGs for gene mutations and clastogenicity/aneugenicity (OECD TGs 476, 487, 473) that could require modifications (Doak et al., 2012). Other tests that are not under OECD TGs are the γ-H2AX and comet assays. The γ-H2AX assay measures DNA double-strand breaks (DSBs) in cells and is based on the phosphorylation of the DNA-associated histone protein H_2AX involved in the repair of DNA DSBs. The in vitro comet assay can provide relevant information when measuring the potential genotoxicity and mechanisms of action of NMs and it has been extensively performed to investigate the genotoxicity of TiO_2-NMs (Landsiedel et al., 2022a). In this perspective, the development of New Approach Methodologies (NAMs), including in vitro *models* such as the 3D reconstructed skin micronucleus (RSMN) assay or co-culture systems, and in silico computational approaches, provide valuable insights into the mechanisms of NM-induced genotoxicity. Whether used individually or in combination, these methodologies enhance safety assessment by generating biologically relevant data on NM interactions and potential hazards. By improving mechanistic understanding, NAMs could offer strong support for regulatory decision-making while reducing reliance on animal testing, an approach that is rapidly gaining momentum as promising methodologies continue to emerge (Doak et al., 2023b; Sewell et al., 2024).

The genotoxic response by NMs is dependent on the cell line used (Landsiedel et al., 2022a; Landsiedel et al., 2022b). TiO_2-NMs have been reported to induce genotoxic effects in lung, liver, and brain cells (Landsiedel et al., 2022a; Brandão et al., 2020a). For example, Charles, Jomini (Charles et al., 2018) demonstrated DNA damage in human liver cells across a range of applications using multiple assays. While most in vitro studies report positive genotoxic effects, exceptions exist; 30 % of comet assay studies (7/24) and 25 % of micronucleus test studies (4/16) found no increase in genotoxicity (Chen et al., 2014). Variations in exposure conditions, experimental designs, and the pchem properties of TiO_2-NMs complicate cross-study comparisons (Sayes et al., 2007). A comprehensive review by Kirkland, Aardema (Kirkland et al., 2022) evaluated 34 datasets from comet and micronucleus assays, concluding that TiO_2-NMs did not exhibit genotoxic effects when studies adhered to OECD recommended methods and utilized well-characterized TiO_2-NM preparations. These findings underscore the critical need for standardized methods and detailed material characterization in genotoxicity assessments. Conflicting results in the literature can be attributed to several factors, including variability in experimental protocols (Stone et al., 2009). While significant standardization efforts have been made in recent years particularly through EU-funded projects aimed at harmonizing pchem characterization and toxicity assessment methods, many studies still exhibit inconsistencies in methodology. This is often due to a lack of adherence to existing test guidelines rather than an absence of standardized protocols. Ensuring compliance with standardized methods will be essential to improve reproducibility and comparability in NM genotoxicity research. Other factors include, (i) the diversity of NMs' pchem properties and insufficient characterization data reporting, and (ii) a limited understanding of the biological effects of TiO_2-NMs, which are linked to their unique pchem characteristics (Sajid et al., 2015) and also to specific response of the in vitro model used (i.e. different responses from different cell lines). Toxicological studies which typically involve exposing cells or organisms to TiO_2-NMs and then evaluating their genotoxic potential, are time-consuming, costly, and require specialized equipment and expertise. Furthermore, animal studies are associated with ethical concerns.

Considering the complexity of NMs, identifying which pchem properties are most relevant to genotoxicity remains an ongoing challenge. While many pchem characteristics are known to influence biological effects, the full extent of their impact and interdependencies is not yet fully understood. Additionally, the wide range of experimental variables influencing genotoxicity outcomes further complicates efforts to establish clear relationships. These limitations highlight the need for *in-silico* NAM approaches, to predict genotoxicity potential while reducing reliance on traditional experimental methods. A widely used *in-silico* approach in nanosafety is quantitative structure–activity relationship (QSAR) modeling (also referred to as nano-QSAR, QNAR, or QNTR). QSAR models use statistical and/or Machine Learning (ML) techniques to establish relationships between NMs pchem properties and their biological activity, including genotoxicity. ML algorithms can be trained on datasets comprising pchem properties, experimental conditions, and genotoxicity target outputs enabling predictive assessments. These models have demonstrated effectiveness, as tools for hazard assessment and prioritization of experimental testing, for example i) Regonia,

[2] https://www.ema.europa.eu/en/documents/report/final-feedback-european-medicine-agency-ema-eu-commission-request-evaluate-impact-removal-titanium-dioxide-list-authorised-food-additives-medicinal-products_en.pdf

[3] https://curia.europa.eu/jcms/upload/docs/application/pdf/2022-11/cp220190en.pdf

Olorocisimo (Regonia et al., 2022) used ML to predict the toxicity potential (log-transformed EC50 values) of 34 modified multi-metallic alloy TiO_2-NMs on Chinese hamster ovary cells. Using density functional theory (DFT)-calculated descriptors and empirical properties (e.g., covalent radius, electronegativity, ionization potential), their random forest (RF) model achieved a high predictive accuracy with an R^2 value of 0.95, despite the small dataset. Given the scarcity of nanotoxicity data, developing models capable of effectively handling small datasets such as trees and regression models remains essential (Furxhi et al., 2020a; Furxhi et al., 2019). However, it is important to acknowledge the risks associated with small datasets, as high predictive accuracy does not necessarily guarantee generalizability. Nevertheless, even the most sophisticated algorithm trained on limited data may be outperformed by simpler models trained on larger, more comprehensive datasets; ii) Trinh, Seo (Trinh et al., 2022) used QSAR-based ML models to assess the mixture toxicity (EC50 and immobilization) of heavy metals combined with TiO_2-NMs using 76 data points with *Daphnia magna*. Descriptors included quantum mechanical properties (e.g., heat of formation, electronic energy, ionization potential, HOMO, LUMO), pchem properties (e.g., particle size, zeta potential), and experimental conditions (e.g., concentration, exposure duration). The RF model performed best achieving an R^2 value of 0.93 for EC50; iii) Sang, Wang (Sang et al., 2022) predicted the mixture toxicity of TiO_2-NMs and heavy metals on human renal proximal tubule epithelial (HK−2) cells using QSAR-based ML models with cell survival rate as the target output. Their dataset of 72 samples incorporated descriptors such as ionization potential, electron affinity, absolute electronegativity, absolute hardness, adsorption energy, and molecular energy. The RF model yielded an R^2 value of 0.95; iv) Likewise, Yuan, Wang (Yuan et al., 2021) developed QNAR models to predict the cytotoxicity of mixtures of TiO_2-NMs and heavy metals on HK-2 cells using a dataset of 36 samples. Mixture descriptors for each quantum mechanical component were calculated and combined using an additive scheme. The RF model achieved the highest R^2 value of 0.87; v) Lamon, Asturiol (Lamon et al., 2018) applied a version of the European Chemicals Agency (ECHA) grouping and read-across workflow to curated publicly available data for six TiO_2 nanoforms (or nanomaterials, anatase and rutile), each characterized by over 100 pchem properties, including crystalline form, shape, aspect ratio, particle size, particle size distributions (water and media), zeta potential and polydispersity (media), isoelectric point, density, surface area and bio-durability, selecting the in vitro comet assay as the target output. Through hierarchical clustering, principal component analysis, and RF models, the study identified two distinct groups based on pchem characteristics. Key challenges in implementing read-across for NMs were highlighted, including experimental variability, the absence of standardized measurement protocols, and the limited mechanistic understanding of NM-induced genotoxicity. These findings underscore the potential of read-across approaches for regulatory decision-making while emphasizing the need for further refinement. In the dataset (provided at the end of the report), the NMs are listed as columns while pchem characteristics are listed as rows, which poses a challenge in terms of ML practices; Models fitted to this type of dataset will be overfitted;

Although these studies demonstrate the promise of in silico models, most have primarily focused on cytotoxicity endpoints (Furxhi et al., 2020b) which are generally more straightforward to predict than genotoxicity. Unlike cytotoxicity, genotoxic responses can vary significantly depending on assay type (e.g., alkaline vs. enzyme-modified comet assay), scoring parameters (e.g., % tail DNA vs. tail moment), and experimental conditions. Additionally, manual data extraction from graphical results presents a major bottleneck, introducing significant time constraints and inefficiencies, posing challenges for large-scale data extraction and modeling efforts. These complexities highlight the need for data accessibility. While several studies have compiled datasets related to TiO_2-NMs, none have systematically focused on modeling their genotoxicity potential while integrating extensive data on pchem properties, experimental conditions, and exposure parameters. This study aims to support nano-informatics by presenting a predictive read-across model for predicting the in vitro genotoxicity potential of TiO_2-NMs. Specifically, this study has: (i) systematically compiled relevant data from the literature, (ii) identified critical data gaps when reporting genotoxicity assessments, (iii) demonstrated the feasibility of modeling approaches for genotoxicity prediction, and (iv) highlighted key variables that significantly influence prediction accuracy.

2. Materials and methods

Fig. 1 snapshots the workflow of this study. The process began with a literature review of in vitro genotoxicity studies involving TiO_2-NMs, specifically using the alkaline and enzyme-modified comet assays. Relevant data were extracted across a range of parameters. Following data cleansing and exploration, the dataset was divided into two subsets to address distinct modeling objectives. Each subset underwent a data preprocessing phase, after which models were trained, validated, and assessed. Key steps included deriving feature importance and defining the models' applicability domain.

2.1. Data extraction

A literature search was conducted for the period 2013–2024 using Google Scholar and PubMed. The search strategy employed Boolean operators to combine keywords, ensuring a comprehensive retrieval of relevant studies. Specifically, the search included combinations such as: ("comet assay" OR "comet") AND ("genotoxicity" OR "DNA damage") AND ("TiO_2" OR "titanium dioxide") AND ("nanomaterials" OR "nanoparticles" OR "nanoform"). This approach ensured that studies mentioning any of the comet assay terms, combined with genotoxicity or DNA damage and TiO_2-NMs, were included. Additional filters were applied to exclude irrelevant studies based on the following criteria: review articles; studies lacking comet assay data; in vitro studies conducted in non-mammalian cells; in vivo studies. Articles specifically investigating the genotoxicity of TiO_2-NMs using alkaline and enzyme-modified comet assays were selected.

-Output target: The target was the numerical values of DNA damage expressed as percentage of DNA in comet tail (% tail DNA), a parameter directly proportional to the frequency of DNA lesions (Lu et al., 2017). Among the descriptors provided by image analysis systems (e.g., "Tail moment," "Tail length," "Olive tail moment," "% tail DNA"), % tail DNA is widely regarded as the most used, reliable and informative metric for quantifying DNA damage (Andreoli et al., 2018; Bossa et al., 2021a; Fatima and Yadam, 2023). Data were gathered from studies employing the standard alkaline comet assay, which detects primary DNA lesions such as single-strand breaks (SSBs), DSBs, and alkali-labile sites. To capture oxidative DNA damage, enzyme-modified versions of the alkaline assay were also included, specifically those incorporating Fpg, OGG1 and EndoIII enzymes. These enzymes enable the detection of specific oxidative lesions: Fpg recognizes oxidized purines, OGG1 detects 8-oxoguanine and other oxidized guanines, and EndoIII identifies oxidized pyrimidines (Armand et al., 2016; Elje et al., 2020). Incorporating enzyme-based assays extends the analysis, allowing the identification of both primary and oxidative DNA lesions. These assays distinguish direct strand breaks (SBs), SBs with oxidative damage (SBs + enzyme), and oxidative damage alone (Net Fpg), calculated as the difference between SBs and SBs + Fpg (Chen et al., 2022; Muruzabal et al., 2021). A dedicated column in the dataset differentiates assay types (alkaline, Fpg, OGG1, EndoIII, Net Fpg), enabling the model to account for distinct DNA damage profiles. This ensures consistency while capturing assay-specific variations in outcomes, enhancing the model's predictive reliability. Numerical values from figures were digitized

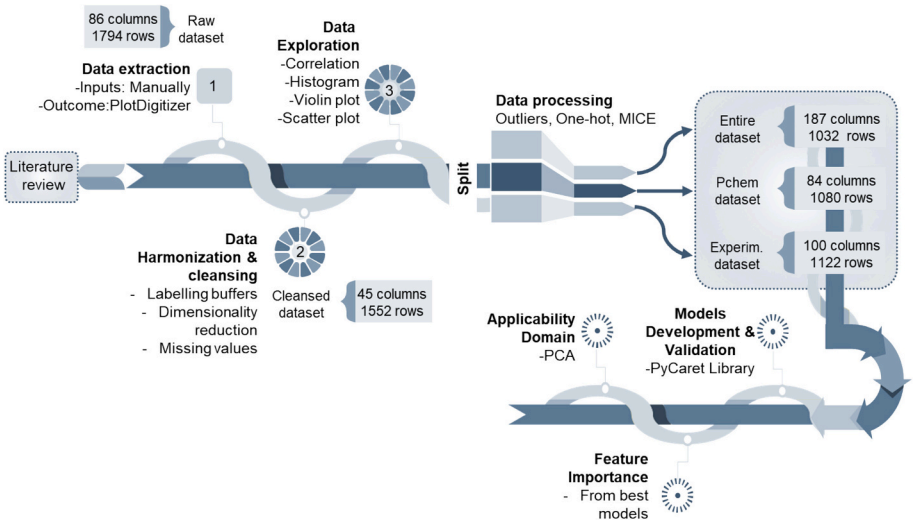


Fig. 1. Workflow of dataset development, processing, and modeling. The process begins with a literature review. Sequential data cleaning and processing steps are indicated (1, 2, and 3). The refined data is subsequently split into three subsets: the Entire, Pchem dataset and Experimental design dataset, for model development and further analysis.

using PlotDigitizer[4] for inclusion in the dataset.

-Inputs: The comet assay is a widely used method; however, inter-laboratory variations in reported damage level exist (Collins et al., 2023). These discrepancies are often attributed to protocol differences. While no validated OECD guidelines exist for the in vitro comet assay, the *Consensus Statement for the Minimum Information for Reporting Comet Assay* (MIRCA) provides recommendations for replicating experiments, distinguishing between "essential" and "desirable" information (Møller et al., 2020). To improve data comparability, facilitate in silico model development, and provide a comprehensive overview of experimental conditions, we extracted detailed information related to: NMs identity, pchem properties and characterization methods, sonication procedures, cellular conditions, enzyme treatment, cold lysis conditions, electrophoresis conditions, statistical and imaging information, cell seeding and preparation, assay and treatment details and exposure conditions.

2.2. Data harmonization & cleansing

For data harmonization the following procedures were followed: i) ***Labelling of buffers:*** A systematic labelling scheme to codify the composition of all buffers including enzyme buffers, cold lysis buffers, and DNA alkaline unwinding buffers was employed. This approach aimed to differentiate buffers based on their chemical composition, enabling the capture of subtle variations in their effects on the target output. Each buffer was codified by categorizing its key components such as salts, chelators, buffers, detergents, and additives where relevant, assigning numerical values to each category, reflecting both the presence and the concentration of these components. ***ii) Dimensionality reduction:*** for categorical features exhibited numerous unique values, a feature merging to reduce dimensionality and improve the interpretability of data, was implemented (Reddy et al., 2020). Retaining numerous unique categories introduces unnecessary complexity that can hinder modeling performance, especially in small datasets. To address this, i) the top 10 most frequent categories for each feature were retained, while less common categories grouped into an "*Other*" category, ii) feature simplification was applied by merging logically related categories into broader representative groups.

For data cleansing: i) rows where the exposure concentration was zero representing cases without NM treatment (blank), were excluded, as they do not provide meaningful information for predicting DNA damage induced by NMs. ii) The distribution of missing data across features was visualized through a heatmap to identify features requiring imputation or removal. A thresholding strategy was applied retaining only features with <30 % missing values, with imputation performed at a later stage to address data gaps. This approach minimized unnecessary data loss while preserving dataset integrity.

2.3. Data exploration

An exploratory data analysis was conducted to investigate the relationships between features. A correlation matrix using Pearson coefficient was generated to examine linear dependencies among numerical features, with coefficients ranging from −1 to 1 (Schober et al., 2018), where 1 indicates a perfect positive linear relationship and − 1 a perfect negative linear relationship. To complement this, various plot types were employed based on their suitability for specific features. i) Histograms to visualize the distribution of numerical features, providing their spread and skewness; ii) Scatter plots to explore relationships between numerical features and the target output as they reveal patterns; iii) Violin plots to display the distribution of DNA damage across categorical features, to show variability and frequency.

2.4. Data processing - pre model

Before training the algorithms, the dataset was split into three subsets. The *entire* dataset includes all features offering a complete overview of the experimental workflow, from the NMs pchem properties, to in vitro procedures, electrophoresis processes and data analysis steps. The objective is to build models that capture the multi-dimensional relationships between materials, experimental design, and biological outcomes. This dataset serves as a reference point for evaluating model performance that target thematically focused subsets designed to derive insights. The separation into two subsets enables a granular exploration of both material properties and electrophoresis process to uncover features driving predictive performance: 1) ***pchem*** dataset includes pchem properties of NMs and variables such as sonication conditions, in vitro characteristics, assay-related features, and exposure conditions (e.g., concentration and duration) are included since they directly or indirectly influence the interpretation of pchem properties in experimental contexts. 2) ***Experimental design*** dataset includes variables related to experimental conditions, particularly: i) cold lysis and electrophoresis

[4] https://plotdigitizer.com/

parameters, ii) statistical results, image analysis software, comets count, and replications, and iii) assay type, positive controls, exposure concentration and duration. To enable differentiation amongTiO_2-NMs we kept the NMs type column. In contrast, the pchem dataset captures NM type indirectly through pchem properties.

Then, the datasets underwent a data processing phase to ensure suitability for ML. This step involved: i) ***handling outliers*****:** a percentile-based outlier removal method was applied to numerical features with non-normal distributions to mitigate the influence of extreme values on results (Smiti, 2020). Features with a skewness greater than 3.0 were flagged for potential outliers. For these features, the 2nd and 98th percentiles were used as lower and upper bounds, respectively, to remove outliers. This approach strikes a balance between removing anomalies and preserving the integrity of the data, ensuring that we retained most of the feature's values. ii) ***One hot-encoding*****:** applied to categorical features in each dataset to convert them into a binary format (Yu et al., 2022), creating a separate column for each unique category within a feature. For example, if a feature contains three unique categories (A, B, and C), one-hot encoding generates three binary columns (A, B, and C) with a value of 1 indicating the presence of that category for a given row, while a value of 0 indicating its absence. ***iii) Missing values imputation*****:** addressed using Multiple Imputation by Chained Equations (MICE), a robust technique for handling incomplete data (Khan and Hoque, 2020). MICE imputes missing values by considering the relationships between variables, offering a more reliable approach than simpler imputation methods (Mera-Gaona et al., 2021). This process was configured to run for up to 10 iterations, ensuring convergence while maintaining computational efficiency.

2.5. Model development & validation

Before modeling, numerical features were normalized with z-score to standardize data. Datasets were split to training and test sets using a random state sampling technique allocating 80 % of data to training set and the remaining 20 % to the test set to evaluate the models' predictive performance. A diverse set of regression models was trained using PyCaret v3.3.2 (an open-source ML library in Python), including CatBoost, Random Forest, Extra Trees, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Gradient Boosting, Decision Tree, K-Nearest Neighbors, Bayesian Ridge, Ridge Regression, Huber, AdaBoost, Elastic Net, and others. Hyperparameter optimization was performed using the *tune model* function in PyCaret (Kim et al., 2024), which applies a Randomized Search algorithm to explore the parameter space and identify the best configurations. The process focused on maximizing performance measured by R^2 metric, leveraging cross-validation for robustness. The models were compared and the top five best-performing models are reported in this study. To assess model performance and generalizability, we employed 10-fold cross-validation that divides the datasets into 10 subsets, training the models on 9 folds while testing on the remaining fold, and repeating this process 10 times so that each fold serves as the test set once. This method provides a reliable performance estimate by averaging results across iterations, minimizing the likelihood of overfitting to a single training-test split. Three evaluation metrics were selected, i) Mean Absolute Error (MAE) that measures the average magnitude of the errors between predicted and actual values, offering a measure of accuracy (Cort and Kenji, 2005; de Myttenaere et al., 2016), ii) Mean Square Error (MSE) which calculates the average of the squared differences between predicted and actual values (Ren et al., 2022) and iii) Coefficient of Determination (R^2) that represents the proportion of variance in the actual values that the model can explain, with values closer to 1 indicating better explanatory power (Chicco et al., 2021). By reporting these metrics, we ensured a comprehensive evaluation of models' accuracy and generalizability. We generated four key plots: i) scatter plot that visualizes the relationship between the actual DNA damage and predicted values; ii) residuals plot, where the residuals (differences between the actual and predicted values) are plotted against the predicted values; iii) residuals distribution (histogram) that shows the distribution of residuals, ideally following a normal distribution around zero; and iv) Quantile-Quantile (Q-Q) plot that compares the distribution of residuals to a normal distribution.

2.6. Feature importance

Feature importance for the *Entire* and *Pchem* datasets was evaluated using the Extra Trees Regressor, based on the final best tuned models (**Section 4.4 & 4.5,** respectively). Extra Trees employ Gini Importance metric to assess a feature's contribution to data splitting within decision trees (Abubaker et al., 2024). This metric is based on the reduction of impurity in child nodes when a feature is used for a split. The overall importance is calculated as the weighted average of impurity reductions across all trees in the forest, with scores normalized to sum to 1. This approach identifies features that significantly reduce uncertainty, revealing those with the highest predictive power. For the *experimental design* dataset, feature importance was calculated using XGB Regressor which also evaluates importance through decision tree-based methods. XGB calculates importance by combining two metrics, the frequency with which a feature is selected for splitting and the magnitude of impurity reduction achieved in each tree (Chen and Guestrin, 2025).

2.7. Applicability domain

Various methods exist for defining the Applicability Domain (AD) of models including leverage- and distance-based approaches such as Euclidean or Mahalanobis distance (Furxhi et al., 2020a). In this study, Principal Component Analysis (PCA) was employed for AD visualization (Wang and Chen, 2023; Kurita, 2019). PCA reduces data dimensionality, projecting it into two dimensions, which facilitates the observation of clusters and identification of outliers that fall outside the model's prediction space. The Euclidean distance-based method was used to compute pairwise distances between training data points in the PCA-transformed space. The 95th percentile of these distances was chosen as the AD threshold, representing the boundary within which predictions are considered reliable. To assess the reliability of predictions for test samples, a scatter plot was generated, visualizing the data points in the PCA-transformed space.

3. Results

3.1. Data extraction

Based on two search engines, 134 articles were extracted of which 61 were deemed relevant for inclusion. The % DNA tail was extracted from figures either in the main text or supplementary materials using PlotDigitizer. The dataset was structured so that each row represents a unique combination of input features based on the experimental design. This approach ensures that even slight variations in input features result in a new row, capturing the full range of experimental conditions. This one-to-one mapping between input feature combinations and their corresponding outputs adheres to a relational database logic, where each row serves as an atomic representation of a unique experiment. The main features that most frequently defined a new row included: concentration, exposure duration, sonication method, cell line, TiO_2 type, and assay type (e.g., comet alkaline or modified assay). Table 1 provides an overview of the input parameters extracted from each study.

The resulting raw dataset comprised 1794 rows and 86 columns containing all extracted information (see *Supplementary Excel: Raw Dataset tab*).

3.2. Data harmonization & cleansing

For data harmonization, the buffer-related features were labelled. To

Table 1
Overview of input parameters extracted from in vitro comet assay genotoxicity studies. The table categorizes key inputs and their associated features to ensure consistency and comparability in data extraction from the literature.

Input categories	Input features
Bibliographic information	Author(s), study title, and Digital Object Identifier (DOI)
Nanomaterial / (nanoparticle / nanoform)	Common names (e.g., Aeroxide P25) and standardized identifiers (e.g., JRC codes NM100, NM101), CAS number, supplier name, catalogue/ batch number, and synthesis method (if synthesized de novo).
Pchem properties and characterization methods	Coating, morphology, crystallinity, particle core size (nm), hydrodynamic size (nm), specific surface area (m^2/g), zeta potential (mV), polydispersity index (PDI), and purity (%). Characterization techniques (e.g., SEM, TEM, DLS). Hydrodynamic size, polydispersity index (PdI), and zeta potential noted with the measurement medium (e.g., water, cell culture medium).
Notes on pchem: For each row, pchem measurements (e.g., hydrodynamic size, zeta potential) were linked to relevant experimental conditions, such as the medium (e.g., water or culture medium), concentration (e.g., 10 μg/mL or 200 μg/mL), or exposure duration. For instance, hydrodynamic size was reported at multiple time points (e.g., 24 h and 48 h), the measurements were tied to the corresponding time point. Similarly, if specific treatments such as sonication were applied, p-chem property values were aligned with these conditions. This linking ensures that pchem properties are contextually accurate and condition-specific. -Missing pchem data for JRC NMs (e.g., NM100–105) were supplemented from the JRC Report[1,2] which provided details such as coating, crystallinity, primary size, hydrodynamic size, PdI, specific surface area, and purity (all measured in water). -Ghosh, Chakraborty (Ghosh et al., 2013) reported an average hydrodynamic diameter of 6000 nm with a PdI of 1.53. As PdI > 1 is implausible, this value was assumed to be a typographical error and adjusted to 0.53	
Sonication procedures	Medium used (e.g., Milli-Q water, BSA, culture medium). Parameters: frequency (kHz), power (W), amplitude (%), duration (total/pulsed/ continuous mode). Instrumentation details and protocols.
Notes on sonication: Assumptions were made when the operational mode was not explicitly stated. For example, if an article reported "10 min of sonication" without further details, it was assumed to be continuous mode unless otherwise specified.	
Aggregation, agglomeration	Assessment of aggregation/agglomeration during experiments (binary format)
Cellular conditions	Light conditions (e.g., experiments conducted in darkness to prevent photoactivation of TiO_2). Medium composition (e.g., DMEM, RPMI-1640, BEGM), including concentrations of FBS (%), glutamine (mM), penicillin (U/mL), and streptomycin (mg/mL).
Notes on Cellular conditions: To ensure consistency: i) the presence or absence of antibiotics (e.g., penicillin/streptomycin) was recorded, ii) If antibiotics were not mentioned, a value of 0 was assigned, indicating their absence, iii) if antibiotics were acknowledged but no concentrations provided, the value was recorded as NaN, signaling incomplete information.	
Cell line information	Cell line (e.g., A549, SH-SY5Y, PBMC), type (e.g., neuroblast-like, epithelial, mononuclear), species (e.g., human, monkey, hamster), tissue (e.g., peripheral blood, brain, alveolar) and origin (e.g., immortalized, primary)
Enzyme treatment	For enzyme-modified comet assays, enzyme quantity, incubation time (min), temperature (°C), and buffer composition.
Cold lysis conditions	Incubation time (h), temperature (°C), buffer composition, and low-melting-point agarose concentration (%) in gels.
Electrophoresis conditions	Alkaline buffer composition, DNA unwinding duration (min), and temperature (°C). Electrophoresis parameters: duration (min), voltage (V), current (V/cm), and temperature (°C).
Notes on Electrophoresis: For studies lacking detailed descriptions of electrophoresis parameters, information was sourced from the referenced publications (Singh et al., 1988).	
Statistical and imaging information	Image analysis software, magnification (x), and statistical evaluation parameters.

Table 1 (*continued*)

Input categories	Input features
Cell seeding and preparation	Multiwell plate specifications, cell density per well/mL/cm^2, cells/comets scored per replicate, and staining compounds.
Assay details	Positive/negative controls, replication (duplicate/ triplicate), assay type, and descriptions of the assays.
Exposure conditions	Exposure concentration (μg/mL) and duration (h) of TiO_2 treatement.

[1] https://products.evonik.com/assets/or/ld/AEROXIDE_TiO2_P_25_TDS_EN_EN_TDS_PV_52043891_en_GB_WORLD.pdf
[2] https://publications.jrc.ec.europa.eu/repository/handle/JRC86291

organize the diverse cold lysis buffer compositions, we grouped them into distinct alphabetical categories based on their key components and concentration, including salts (NaCl, NaOH), chelating agents (EDTA, Na_2EDTA), buffer systems (Tris, Tris-HC1, Tris-base), surfactants/detergents (Triton X-100, Sodium sarcosinate, N-lauroylsarcosine 1 %) and additives (DMSO). The same approach was followed for the electrophoresis alkaline buffers (NaOH molarity, EDTA type and concentration, additional components) and the enzyme buffers (HEPES, KCl, EDTA, BSA, pH). Table 2 outlines the step-by-step categorization approach used for buffer labeling.

Dimensionality Reduction: The number of unique shape descriptors was reduced from 20 to 8 while preserving the morphological information by grouping geometrically similar shapes with minor variations under broader representative categories (Table 3).

For crystallinity phase, categories followed by a ratio like "anatase/rutile_25:75" were merged into "anatase/rutile". This simplification prevents the model from treating materials with minor ratio variations as entirely distinct entities, improving generalization. Coating was collapsed into two categories, "coated" and "uncoated" to reduce sparsity while retaining meaningful information about the fundamental distinction between those materials. As uncoated samples account for 78 % of the data, this simplification helps the model focus on significant differences without being influenced by variations in specific coating types present in small sample sizes. Non-human cell species (hamster, monkey, mouse etc) were grouped into a "non-human" category. Rare occurrences of species can hinder model learning, due to insufficient data for reliable predictions. For cell name, cell type, cell tissue, and cell medium the top 10 categories based on their frequency distribution were retained. Less common categories were grouped under "other". The same logic was applied to positive control. This simplification ensures the model focus on well-represented features in the dataset. To categorize the statistical analyses used in the studies for the comet results, methods were grouped based on their underlying assumptions and nature, such as parametric, non-parametric, or regression-based methods. For example, *ANOVA and parametric post-hoc tests* assume data follows a normal distribution. This category includes variations of ANOVA, one-way and two-way, along with parametric post-hoc tests like Tukey, Bonferroni, Dunnett, and Holm-Sidak tests. *ANOVA and non-parametric post-hoc tests* are used when normality assumptions are not met, for example, Kruskal-Wallis rest serves as a non-parametric alternative to ANOVA, often followed by the Mann-Whitney *U* test for pairwise comparisons, Wilcoxon test for paired data and the Kolmogorov-Smirnov test for distribution comparisons. Post-hoc corrections, such as the Bonferroni correction, are applied in these tests for multiple comparisons. *Regression Analysis* group includes linear regression analysis, which may include post-hoc tests or transformations depending on the nature of the data such as Shapiro-Wilk and Bartlett's test and student's *t*-tests.

For data cleansing: i) rows where the exposure concentration was zero and ii) the negative control column, were removed. Negative controls, typically untreated cell cultures, are included in studies to confirm that observed effects in treated samples are attributable to NM exposure rather than external factors. iii) Columns representing methods used to

Table 2
Step-by-Step categorization approach for the buffer's codification.

	Cold lysis buffer						
Label	Buffer Components (pH 10)	Salt	Chelator	Buffer	Detergent	Additive	Labels
A	2.5 M NaCl, 100 mM Na_4EDTA, 10 mM Tris-HCl, NaOH, 1 % Triton X-100	1	2	2	1	0	A_12210
B	2.5 M NaCl, 100 mM Na_2EDTA, 0.4 M Tris-Base, 10 % DMSO, 1 % Triton X-100	1	1	6	1	1	B_11611
C	2.5 M NaCl, 100 mM Na_2EDTA, 10 mM Tris, 1 % sodium sarcosinate, 1 % Triton X-100	1	1	1	1	0	C_11110
D	2.4 M NaCl, 100 mM Na2EDTA, 5 mM NaOH, 10 % DMSO, % Triton X-100	2	1	0	0	0	D_21000
E	2.5 M NaCl, 100 mM EDTA, 10 mM Tris, 1 % Triton X-100	0	3	0	0	0	E_03000
F	2.5 M NaCl, 100 mM EDTA, 10 mM Tris-HCl, 10 % DMSO, 1 % Triton X-100	1	3	2	1	1	F_13211
G	2.5 M NaCl, 100 mM Na_2EDTA, 10 mM Tris, 1 % Triton X-100	1	1	1	0	0	G_11100
H	2.5 M NaCl, 100 mM Na_2EDTA, 10 mM Tris, 10 % DMSO, 1 % Triton X-100	1	1	1	1	1	H_11111
I	2.5 M NaOH, 100 mM EDTA, 10 mM Tris, 1 % Triton X-100,	3	3	0	1	0	I_33010
J	2.5 M NaCl, 100 mM EDTA, 1 % Triton X-100, 1 % N-laurosyl sarcosine, 10 % DMSO	1	3	0	1	1	J_13011
K	2.5 M NaCl, 100 mM Na_2EDTA, 10 mM Tris-base, 1 % Triton X-100, 1 % N-lauroylsarcosine, 10 % DMSO	1	1	3	1	1	K_11311
L	2.5 M NaCl, 100 mM EDTA, 10 mM Tris with 8 g NaOH, 1 % Triton X-100, 10 % DMSO	1	3	0	1	1	L_13011
M	2.5 M NaCl, 100 mM Na_2EDTA, 10 mM Tris–HCl, 1 % Triton X-100, 10 % DMSO	1	1	1	0	1	M_11101
N	2.5 M NaCl, 100 mM EDTA, 10 mM Tris, 10 % DMSO, 1 % Triton X-100	1	3	1	1	1	N_13111
O	2.5 M NaCl, 100 mM EDTA, 10 mM Tris, 10 % DMSO	1	3	1	0	1	O_13101
P	2.5 M NaCl, 100 mM EDTA, 10 mM Tris-HCl, 1 % Triton X-100	1	3	2	1	0	P_13210
Q	2.5 M NaCl, 100 mM Na_2EDTA, 10 mM Tris	1	1	1	0	0	Q_11100
R	2.5 M NaCl, 100 mM Na_2EDTA, 10 mM Tris, 1 % sodium sarcosinate, 10 % DMSO, 1 % Triton X-100	1	1	1	1	0	R_11110
S	2.5 M NaCl, 100 mM Na_2EDTA, 10 mM Tris-base, 10 M NaOH, 1 % Triton X-100,	1	1	3	0	0	S_11300
T	2.5 M NaCl, 100 mM Na_2EDTA, 10 mM Tris–HCl, 1 % Triton X-100	1	1	1	0	0	T_11100
U	2.5 M NaCl, 100 mM Na_2EDTA _8mM Tris-HCl_1 % Triton X-100, 10 % DMSO	1	1	4	0	1	U_11401
V	2.5 M NaCl, 100 mM $Na_2EDTA._2H_2O$, 10 mM Tris–HCl, 1 % Triton X-100, 10 % DMSO	1	1	1	0	1	V_11101
W	0.2 M NaOH, 100 mM EDTA, 0.1 % Triton X-100	4	3	0	1	0	W_43010

	Electrophoresis alkaline solution buffer				
Label	Buffer Components	NaOH	EDTA	Additional	Labels
A	0.3 M NaOH, 1 mM Na_2EDTA	2	1	0	A_210
B	0.3 M NaOH, 10 mM Na_4EDTA	2	5	0	B_250
C	0.2 M NaOH. 1 mM EDTA	1	4	0	C_140
D	0.2 M NaOH, 1 mM EDTA, 0.1 % Triton-X	1	4	1	D_141
E	0.3 M NaOH, 1 mM Na_2EDTA, 0.2 % DMSO	2	1	2	E_212
F	0.3 M NaOH, 1 mM EDTA	2	4	0	F_240
G	0.3 M NaOH, 1 mM EDTA, 10 % DMSO	2	4	3	G_243
H	0.3 M NaOH, 1 mM $Na_2EDTA.2H20$	2	1	0	H_210
I	5 M NaOH, 200 mM Na_2EDTA	3	2	0	I_320
J	10 M NaOH, 200 mM EDTA	4	3	0	J_430

	Enzyme treatment solution buffer						
Label	Buffer Components	HEPES	KCl	EDTA	BSA	pH	Labels
A	10 mM HEPES, 0.1 M KCl, 10 mM Na_2EDTA, 0.1 mg/mL BSA, pH 7.4	1	1	1	1	1	A_11111
B	10 mM HEPES, 0.1 M KCl, 0.5 mM EDTA,0.2 mg/mL BSA, pH 8.0	1	1	0	0	2	B_11002
C	10 mM HEPES, 0.1 M KCl, 10 mM Na_2EDTA, pH 7.4	1	1	0	0	1	C_11001
D	40 mM HEPES, 0.1 M KCl, 0.5 mM EDTA, 0.2 mg/mL BSA, pH 8.0	0	1	2	2	2	D_01222
E	40 mM HEPES-KOH, 0.1 M KCl, 0.5 mM EDTA, 0.2 mg/mL BSA, pH 8.0	3	1	2	2	2	E_31222
F	40 mM HEPES-KOH, 0.1 M KCl, 0.5 mM Na_2EDTA, 0,2 mg/mL BSA, pH 8.0	3	1	3	0	2	F_31302
G	40 mM HEPES,0.1 M KCl,0.2 mM EDTA, 0.2 mg/mL BSA, pH 8.0	2	1	4	2	2	G_21422
H	40 mM HEPES, 0.1 M KCl,0.5 mM Na_2EDTA, 0.2 mg/mL BSA, pH 8.0	2	1	3	2	2	H_21322

characterize pchem properties were removed, as they were not directly relevant to the primary analysis. iv) Of the 61 studies, 19 incorporated enzymes in their experimental design, contributing 557 rows to the dataset. However, due to limited data availability on enzyme buffer parameters, enzyme-specific details were excluded from the modeling process. Columns removed included *quantity of enzyme used*, *incubation time with enzyme (*min*)*, *incubation temperature with enzyme (°C)*, and *enzyme buffer label*. Despite this exclusion, the *Assay type* feature was retained, to capture distinctions between the alkaline comet assay and the enzyme-modified version. This ensured that the assay context including enzyme specificity, was still represented in the modeling process, albeit without detailed enzyme-related information.

Missing values handling: A heatmap was generated with Fig. 2 visualizing the distribution of missing values across features, providing a data completeness picture.

A strategy for handling missing data was applied by setting a threshold: columns with <30 % missing data were retained, and imputation was later employed while those exceeding this threshold were removed. Approximately, half of columns were removed. Example of removed inputs include DNA unwinding temperature (67.5 %), PdI measured in water (48 %) and in cell medium (68.61 %), zeta potential measured in water or stock suspension (59 %) and in cell medium (62.6 %), sonication-related features such as power (52 %), protocol (56 %), temperature (57 %), amplitude (69.9 %) and frequency (82.2 %), light conditions (52 %), electrophoresis-related features such as temperature (59 %) and voltage (V) (36.6 %). However, voltage expressed in V/cm was retained as it had <30 % missing values. Several cell density-related features were removed, including density per well (64.4 %), per mL (66.9 %), per slide (80.5 %), and per cm^2 (90 %). Lastly, magnification used for imaging had 74.7 % missing data and was therefore excluded.

Table 3
Harmonized shape categories and their corresponding unique shape descriptors.

Harmonized Category	Unique names
Spherical/ Quasi - Spherical	Unique categories: Spherical, Quasi-spherical, Circular slightly elongated, Elliptical/ spherical, Spherical/ellipsoidal, or cuboidal. These shapes are all rounded, with slight variations in elongation but retaining a primarily spherical or near-spherical nature.
Ellipsoidal/ Elongated	Unique categories: Elongated, Rounded/ellipsoidal, Irregular/ ellipsoidal, elongated. These shapes describe elongated structures with a rounded or ellipsoidal geometry.
Rod-like	Unique categories Rods, Spherical/rod, Wire. These shapes describe elongated structures with defined axial symmetry, where one dimension is longer than the others.
Irregular	Unique categories: Irregular laminar, Irregular-tetragonal, Irregular/semi-spherical, spherical/irregular/rod, Irregular euhedral. These shapes describe particles that lack a clear geometric form and show irregularities.
Polyhedral	Unique categories: Polyhedral, Bipyramids. These shapes describe particles with well-defined, multi-faceted surfaces, distinct from irregular shapes.
Plate-like	Platelets have a distinct, flat, and thin morphology, making them a separate category.
Hexagonal	Hexagonal particles have a distinct geometry making, them a separate category

Following data cleansing and the removal of duplicates, a total of 45 features remained with 1551 rows for analysis *(see Supplementary excel: Dataset v00 tab)*. Table 4 presents the final set of features retained in the entire dataset. For categorical features, the table lists the top occurrences based on frequency. For numeric features, the mean, minimum, and maximum values are presented. The table also reports the final percentage of missing values for each feature, providing an overview of final data completeness.

3.3. Data exploration

To explore the relationships between variables, a Pearson correlation matrix was generated (Fig. 3). A negative correlation of core diameter with surface area (−0.27), which aligns with fundamental particle physics principles. However, no correlation was found between core diameter and the target output. Hydrodynamic sizes measured in water and medium showed a moderate positive correlation (0.45), suggesting the sizes are fairly connected between the two environments. The hydrodynamic size in the medium has a slightly stronger correlation with DNA damage (0.17) than in water (0.06) suggesting that size in a biological context may be more relevant to DNA damage. Furthermore, correlation between hydrodynamic size in medium and the concentration of streptomycin (0.36) weakly suggest that the presence of streptomycin might influence the aggregation or dispersion of NMs in the medium; The relationship between sonication time and cold lysis time (0.27) potentially reflecting protocol driven relationships rather than causal links (Fig. 3). Similarly, correlation between electrophoresis current and exposure concentration (0.41) or the concentration of fetal bovine serum (FBS) in the cell medium correlated positively with both cold lysis temperature (0.42) could be due to experimental settings adjusted to ensure clear visualization of DNA migration rather than an intrinsic effect of NM concentration on electrophoretic behavior. The relationship between sonication time and DNA damage (0.26) appeared weak. It is important to note that many correlations in the dataset are weak, indicating that DNA damage is likely driven by non-linear interactions rather than single variables.

From Fig. 4, distributions of some numerical features can be observed. Core sizes are primarily concentrated below 50 nm with a tail extending toward larger diameters. Hydrodynamic size measured in water exhibits a right-skewed distribution, with most particles being smaller than 500 nm. However, the presence of outliers in the range of thousands suggests agglomeration in some samples. Similarly, the hydrodynamic size measured in medium is also right-skewed, though the distribution is more spread out compared to water, reflecting greater variability in biological media (Fig. 4, *C)*. Sonication times are

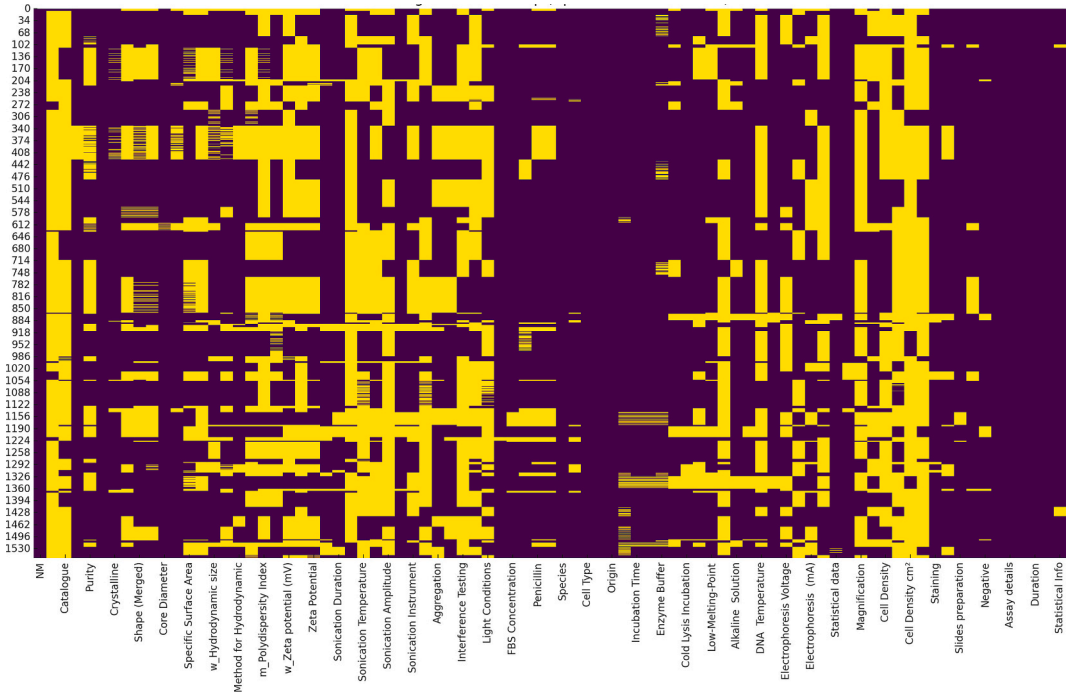


Fig. 2. Heatmap of missing values in the raw dataset. Yellow indicates missing data, while dark areas represent complete data. The y-axis represents dataset row indices, corresponding to individual experimental entries. The x-axis lists the dataset features. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Final set of categorical and numerical features retained in the full dataset after data cleansing with missing values and descriptive statistics.

Categorical features				Numerical features					
	Unique	Missing (%)	Top occurrences examples		Unique	Missing (%)	mean	min	max
NMs Type	11	0	NM100, Aeroxide P25, NM103, etc.,	Core Diameter (nm)	11	0.65	47.4	5	352
Source of NMs	11	0	JRC, Sigma Aldrich, Evonik Degussa, etc.,	Specific S. Area (m^2g)	44	26.6	60	9.2	356
Coating	4	0	Coated, Uncoated	Hydro Size in water (nm)	80	19.6	318.5	23	7289
Crystalline Phase	1	0	anatase, rutile, anatase/rutile, others	Hydro Size in medium (nm)	117	22.4	441.7	23.3	2419
Shape Merged	8	24.1	Spherical / quasi-spherical, ellipsoidal / elongated, irregular, red-like, cube, etc.,	Sonication Time (min)	17	8.65	18.9	1	210
Sonication Medium	6	8.5	BSA, water, medium, PBS, stock	FBS Concentration (%)	6	3.81	7.9	0	15
Sonication Mode	2	6.6	continuous, pulsed	Glutamine (mM)	4	12.9	2	0	4
Aggregation (Y/N)	2	28.6	yes, no	Penicillin (U/mL)	5	15.16	91.1	0	1000
Culture Medium	11	0	DMEM, RPMI 1640, BEGM, Hams F12, etc.,	Streptomycin (mg/mL)	9	15.2	12.9	0	100
Cell Species	2	0	human, non-human	Cold Lysis Incubation (h)	6	7.8	6.4	0.7	24
Cell Name	11	0	A549, BEAS_2B, Caco2, TK6, others, etc.,	Cold Lysis Incubation (°C)	3	22.5	2.5	0	5
Cell Type	11	0	epithelial, hepatocytes, B-lymphoblasts, fibroblast	LMP Concentration (%)	8	15	0.73	0.5	1.5
Cell Tissue Organ	11	0	alveolar, bronchial, intestine, peripheral blood, liver,	DNA Unwinding Time (min)	4	6.3	25.8	15	60
Cell Origin	3	0	immortalized (cancer-derived, non-cancer), primary	Electrophoresis (min)		4.1	25.9	15	50
Cold Lysis Code	11	0	E_03000, H_111111 P_13210, U_11401 etc.,	Electrophoresis (V/cm)	8	29.5	0.94	0.24	1.33
Electrophoresis Buffer Code	10	11	F_240, A_210, G_243, D_141, H_210, etc.,	Electrophoresis (mA)	5	23.8	299.4	280	400
Statistical Processing	6	1.4	ANOVA, parametric, ANOVA non-parametric, etc.,	Multiwell Plate for genotox.	7	27.8	34.1	6	96
Image Analysis Software	11	0	Comet IV, Komet 5.5, Trevigen Comet Software, CaspLab etc.,	Number of Cells or Comets Scored	3	13.9	73.6	25	100
Staining compound	9	4	ethidium bromide, SYBR-Gold, SYBR Green, DAPI, propidium iodide, etc.,	Concentration of Treatment (μg/mL)	85	0	99.6	0.001	9983
Positive Control	11	0	H_2O_2, MMS, Ro 19–8022, EMS, riboflavin/ UVA, etc.,	Duration of Exposure (h)	26	0	55.2	0.25	1460
Assay	5	0	alkaline comet, Fpg-modifed, net Fpg-modifed, comet-Endo III, OGG1 Modified	% DNA tail	170	0	12.2	0.002	79
Assay Details	7	0	Single-cell gel, High-throughput, Trevigen CometSlide, mini gel etc.,						
Statistical Information	7	2.3	Mean ± SD, mean ± SEM, average ± SD, median ± SD, etc.,						

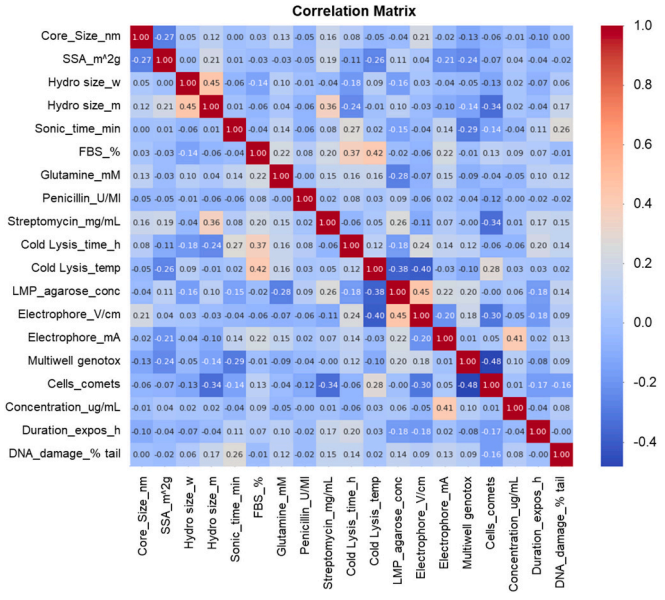


Fig. 3. Pearson correlation matrix showing the linear dependencies among numerical features in the entire dataset.

concentrated within 0–50 min, with relatively few samples undergoing longer periods. Cold lysis duration shows a bimodal distribution, with peaks around 3 and 14 h, reflecting differences in experimental protocols. The concentration of LMP agarose displays a multimodal distribution, with distinct peaks at 0.6 %, 0.8 %, 1.0 %, and 1.2 % (Fig. 4, *D)*. Electrophoresis unwind time exhibits a bimodal pattern, with peaks around 15 and 30 min and some experiments extending to 40 min. Exposure duration predominantly falls under 200 h, with a long tail extending up to 1400 h. This distribution suggests that most experiments are conducted over short durations, with a smaller subset extending to longer exposure times. The percentage of DNA damage exhibits a skewed distribution, with most values clustering between 0 and 30 %. However, a long tail extends toward higher values, reflecting a wide range of genotoxicity outcomes under different experimental conditions (Fig. 4, *B)*. The rest of the histograms can be found in ***Fig. S1***.

The patterns between the numerical features and DNA damage were explored by scatter plots (Fig. 5*)*. **Core diameter** shows a weak negative relationship with DNA damage, suggesting that larger NMs might slightly reduce DNA damage, though pattern is not strong aligning with earlier correlation observations (Fig. 5, *A)*. **Surface area** and **hydrodynamic sizes** exhibit similar weak tendencies, indicating a potential but limited role of particle size-related factors in influencing DNA damage. Exposure concentrations are predominantly below 1000 μg/mL, with a few studies testing higher concentrations. Even after filtering outliers, a slight positive relationship is visible, suggesting a potential dose-response relationship (Fig. 5, *I)*. **Exposure duration** shows a weak positive movement, with higher DNA damage percentages observed at longer durations (200–1400 h) hinting at cumulative effects over time (Fig. 5, *E)*. A weak correlation between **electrophoresis voltage** and DNA damage is observed, indicating that higher voltages might enhance damage detection. For **DNA unwind time**, variability is evident across time points (20, 30, and 40 min), with no clear tendency. The scatter

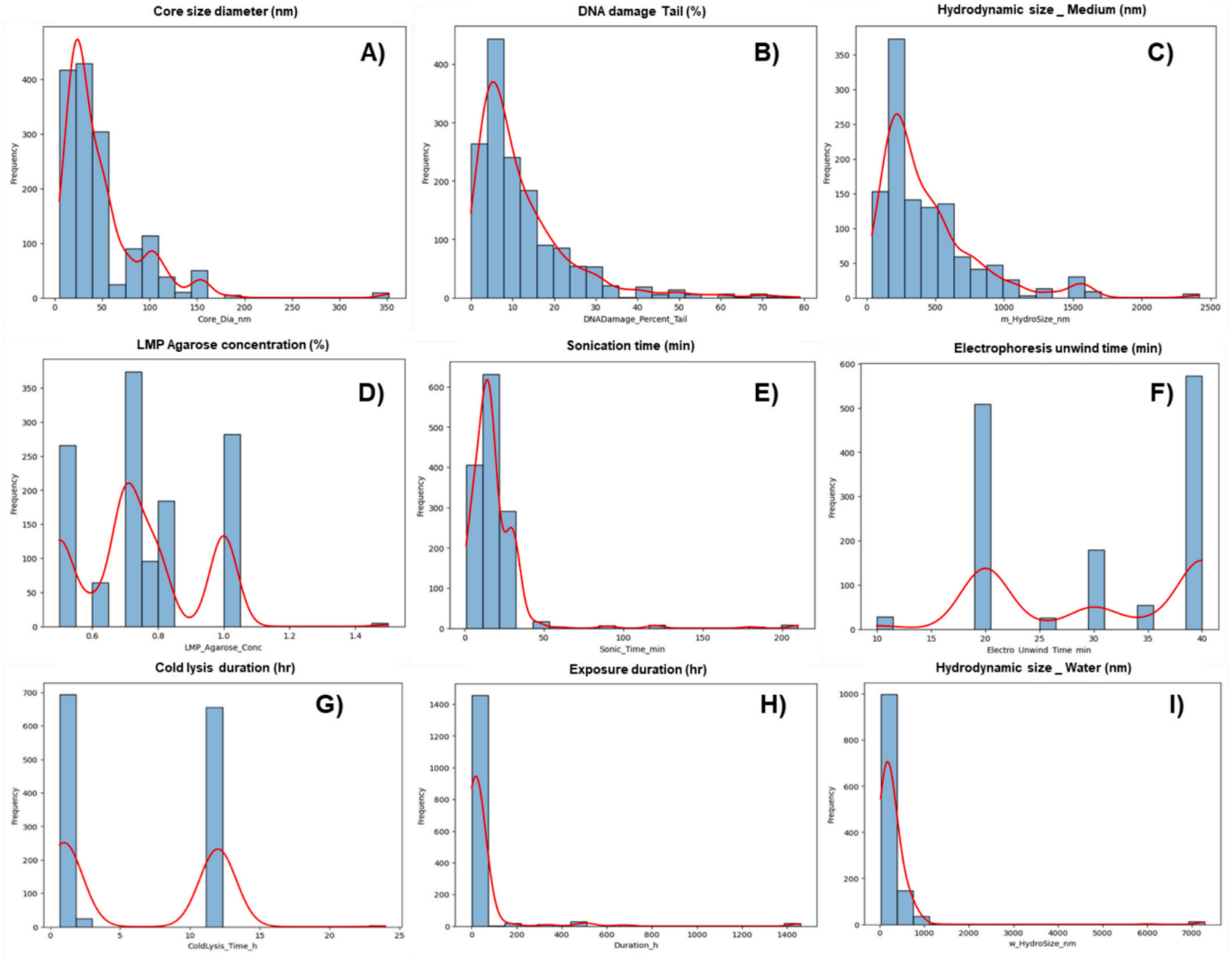


Fig. 4. Histograms examples showing the distribution of numerical features in the entire dataset. (A) Core diameter (nm), (B) DNA damage percentage (% tail DNA), (C) Hydrodynamic size in medium (nm), (D) LMP agarose concentration (%), (E) Sonication time (min), (F) Electrophoresis unwind time (min), (G) Cold lysis time (h), (H) Exposure duration (h), and (I) Hydrodynamic size in water (nm).

plots for **sonication time** do not show distinct patterns, though shorter sonication times (<50 min) appear to be associated with higher damage, with the effect diminishing as sonication time increases (Fig. 5, *G)*. However, some elevated DNA damage is also observed at prolonged sonication times (e.g., 180 min). Furthermore, as most studies in our dataset utilized lower sonication times, the observed effects are more pronounced at shorter durations, while limited data at longer sonication times (1 study) makes it difficult to determine clear patterns. While individual features exhibit weak correlations, the overall picture suggests that DNA damage is likely influenced by multi-factorial interactions rather than single variables. The scatter plots for all features are presented in ***Fig. S2***.

The distribution of the DNA damage across the different categorical features are visualized by mean of violin plots (Fig. 6*)*. Among tissues, testes exhibit higher median values and wider distributions, suggesting greater susceptibility to DNA damage. Embryonic and liver tissues display tighter distributions at the lower end of DNA damage. Tissues like kidney, intestine, alveolar, and bronchial show narrow distributions centered around 0–20 %, reflecting consistent DNA damage levels (Fig. 6*, A)*. Peripheral blood shows a mixed response with a long tail extending to higher DNA damages, indicating variability in susceptibility. Materials from Evonik Degussa show the most variable range of DNA damage, with values extending up to 80 %. Sigma Aldrich and JRC materials display narrower distributions, with most DNA damage clustered between 5 and 30 %. However, both show long tails with instances of higher damage. Sigma Chemical exhibits a compact distribution with most values concentrated at lower values (0–20 %) and minimal occurrences of higher damage (Fig. 6*, B)*. NM synthesized by forced hydrolysis show a highly compact profile with nearly all values below 20 %. NM synthesized by solvothermal processes have a broader distribution, with samples reaching values around 40 %. NM synthesized by laser pyrolysis exhibits wider variability but with overall lower DNA damage values. Irregular, ellipsoidal, spherical, and rod-like shapes show the most variability in DNA damage, with the potential to induce high levels of damage, while hexagonal, cube and polyhedral shapes tend to induce lower and more consistent levels of DNA damage (0–20 % for most cases) (Fig. 6*, C)*. The extent of DNA damage varies across different types of NMs. Aeroxide P25 and NM100 exhibit the greatest variability, with DNA damage percentages ranging from very low to as high as 80 %. In contrast, NM101 - NM104 show narrower distributions of DNA damage, with values predominantly clustered around 20 % (Fig. 6*, D)*. This could indicate more consistent and predictable DNA damage outcomes for these materials. E171 demonstrates a tight distribution centered around 10–15 %, indicating relatively low and consistent levels of DNA damage. The anatase/rutile mixture exhibits the widest range of DNA damage at lower values with a long tail

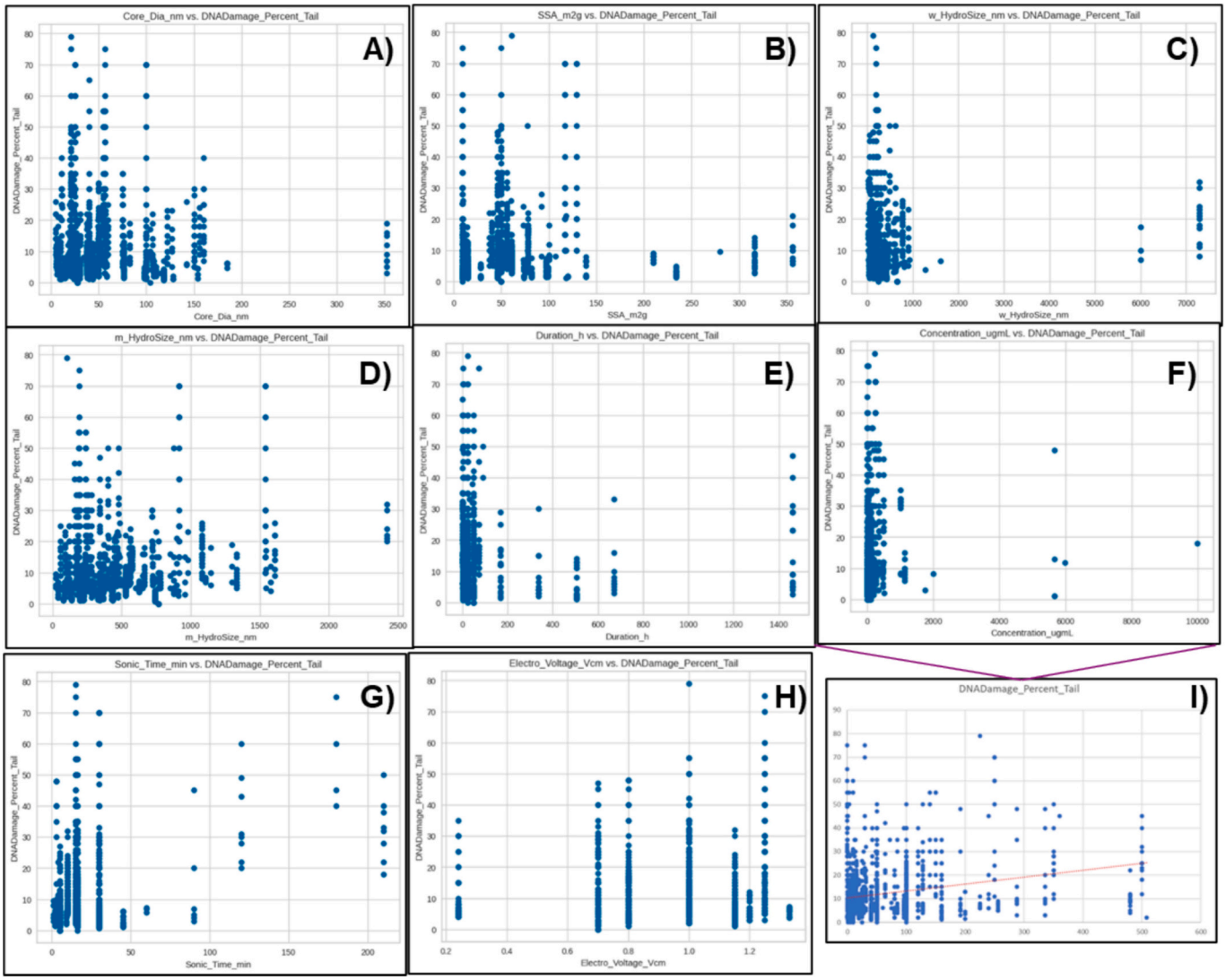


Fig. 5. Examples of numerical features scatter plots with the target output. Scatter plots showing trends between numerical features and % tail DNA. (A) Core diameter, (B) Specific surface area, (C) Hydrodynamic size in water, (D) Hydrodynamic size in medium, (E) Exposure duration, (F) Concentration, (G) Sonication time, (H) Electrophoresis voltage, and (I) Concentration after removal of extreme value.

extending up to 80 % (Fig. 6*, E)*. Anatase alone shows a similar distribution, with a less pronounced profile at higher values. The majority of values cluster between 0 and 20 %, with a median around 10 %, indicating less variability compared to the anatase/rutile phase. Rutile exhibits a broader distribution with a higher frequency of DNA damage between 0 and 30 % and a tail extending to higher values, suggesting greater variability and potentially higher genotoxicity than anatase. These results imply that mixed phases may be more variable and potentially more hazardous than pure anatase or rutile phases. Regarding the cell medium, MEM shows a broad range of high median value of DNA damage, with higher percentiles reaching close to 80 %. RPMI 1640, DMEM and EME shows also a broad range of DNA damage with median percentiles around (0–20 %). Hams F12 and DMEM F12, appear to have more consistent distributions (Fig. 6*, F)*. DMEM/GlutaMAX shows a compact distribution with minimal variation of DNA damage. **Fig. S3** captures violin plots for all features.

3.4. Data processing – pre models (entire dataset)

Outlier analysis resulted in the removal of rows leaving a dataset with 1032 rows comprised of 20 numeric and 24 categorical features, the latter which were expanded through one-hot encoding, increasing the feature space to 187 features. To handle missing data, the MICE approach was applied, and after imputation, the dataset was split into training (722 rows) and test (310 rows) sets enabling the development and evaluation of models.

3.4.1. Model development & validation

Table 5 shows the top-performing models of the entire dataset with *Extra Trees Regressor* outperforming the other algorithms.

The scatter plot (Fig. 7**,** A) of actual vs. predicted values demonstrates strong model performance, with most points aligning along the diagonal. However, deviations become noticeable for values exceeding ~40. The Q-Q plot (Fig. 7**,** B) residuals largely follow the red line, with deviations at the extremes, indicating heavy-tailed behavior. The residual plot (Fig. 7, C**)** indicates residuals are centered around zero but display a slight fan shape, suggesting heteroscedasticity with increasing variance at higher predicted values. The residual distribution (Fig. 7**,** D) approximates normality, though a few outliers on the right suggest the model underestimates actual values for certain data points. Those deviations were expected, as the distribution earlier (Fig. 4) shows that DNA damage exhibits a skewed pattern, with most values clustering between 0 and 30 %, while a long tail extends toward higher values. Since high-damage cases are less frequent in the dataset, fewer training

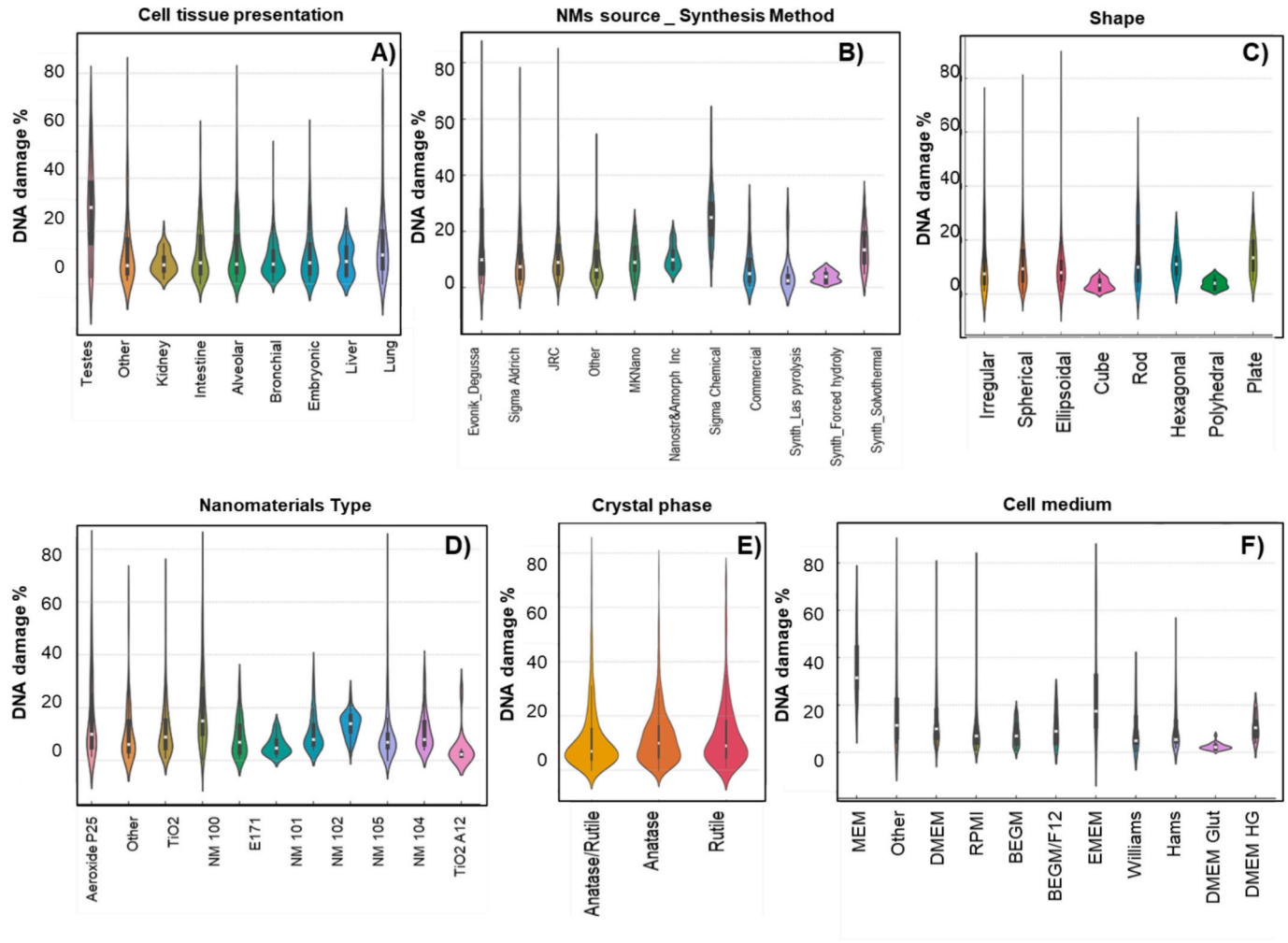


Fig. 6. Examples of violin plots illustrating the distribution of DNA damage percentages across various categorical features. (A) Cell tissue presentation, (B) Nanomaterial source and synthesis method, (C) Nanomaterial shape, (D) Nanomaterial type, (E) Crystal phase, and (F) Cell medium.

Table 5
Top 5 best-performing models on the entire dataset, evaluated using a 10-fold cross-validation approach. The table presents validation metrics, including R^2, Mean Squared Error (MSE), and Mean Absolute Error (MAE), to compare model performance.

Model	R^2	MSE	MAE
CatBoost Regressor	0.813	23.158	2.568
Random Forest Regressor	0.806	24.005	2.664
Extra Trees Regressor	**0.848**	**18.896**	**2.211**
XGB Regressor	0.845	19.246	2.399
Gradient Boosting Regressor	0.658	42.413	3.478

samples exist in this range, leading to prediction variability. As a result, the model's predictive accuracy is reduced in this range, reflecting the inherent data distribution rather than a modeling limitation.

3.4.2. *Feature importance*

The features with the highest influence to the model's predictions is exposure concentration followed by cell culture medium, highlighting the significant role of the surrounding medium in determining the genotoxicity of NMs (Table 6). NMs type ranks third, emphasizing the impact of material-specific properties on the model's outcomes. This is followed by the choice of cell line and the positive controls, Beyond the top five features, cell type also contributes to the predictions, reflecting broader cell classifications and their relevance in genotoxicity assays. The assay type feature captures variations in importance based on the experimental method used to evaluate genotoxicity. Experimental conditions, such as electrophoresis current and exposure duration, also influence DNA damage outcomes. Although less significant, the method of synthesis and/or source of the NMs further contributes to the model's predictions, underscoring its impact on genotoxic effects. The entire results of the feature importance is shown in ***Fig. S4***.

3.4.3. *Applicability domain*

The PCA plot (Fig. 8) shows most training data points clustered around the center, with principal component scores close to zero on both axes, indicating a balanced distribution. The spread of training data across the PCA space reflects significant variability, capturing a broad feature space. Test data points (green) overlap extensively with the training data, suggesting that the test set largely resides within the same applicability domain as the training set, ensuring reliable model evaluation.

3.5. *Data processing – pre models (Pchem and experimental datasets)*

The pchem dataset consists of 1080 rows and 25 features, comprising 12 numeric and 13 categorical features. The categorical features were transformed through one-hot encoding, increasing the dimensionality to 84 features. After pre-processing dataset was split into training (756 rows) and test (324 rows) sets. The experimental dataset consists of 1122 rows and 21 features, of which 10 were numeric and 11

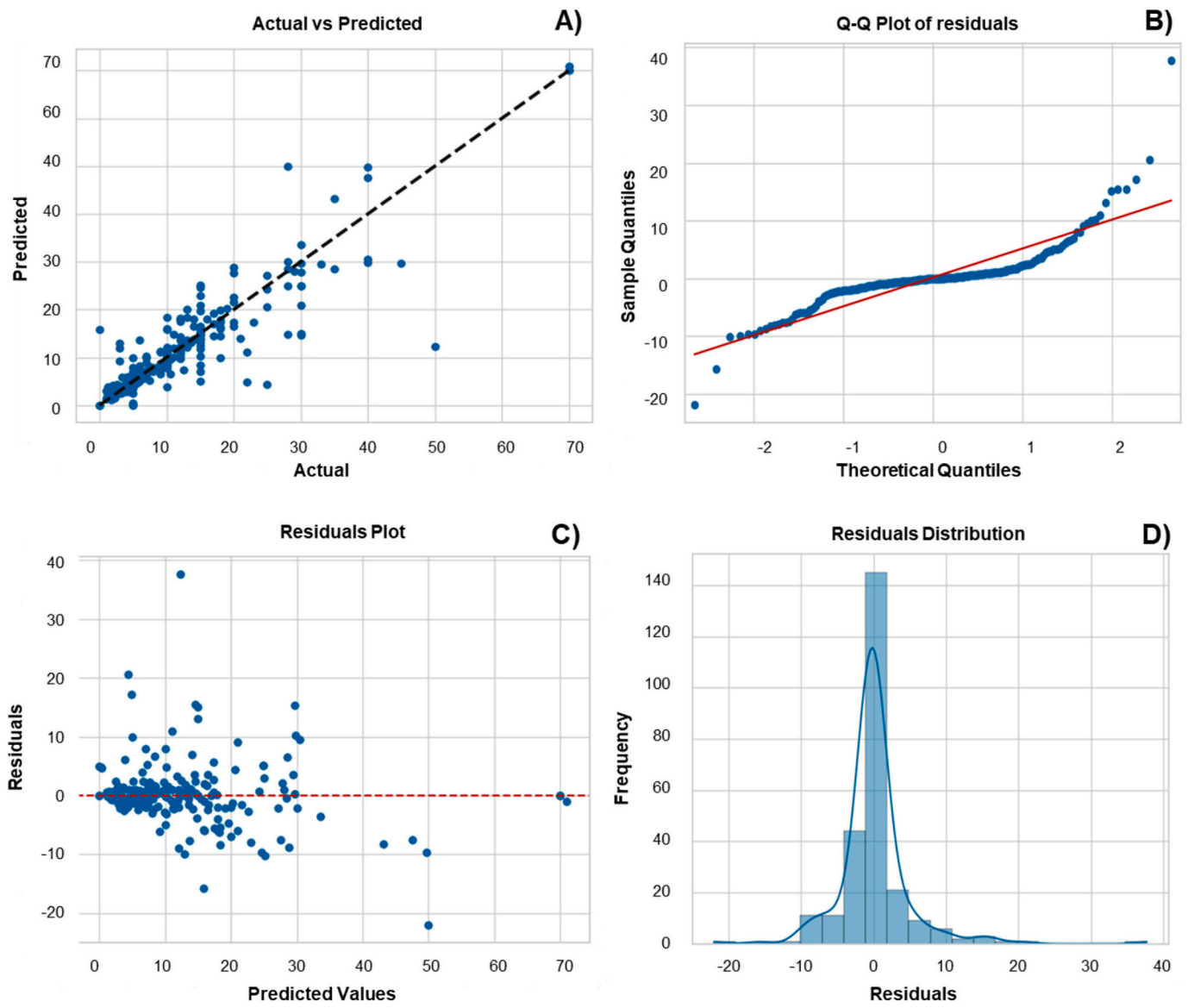


Fig. 7. Performance evaluation plots for the best-performing model (ExtraTreesRegressor). (A) Scatter plot of actual vs. predicted values. (B) Q-Q plot. (C) Residuals plot. (D) Residuals distribution.

categorical. Similar preprocessing steps, including one-hot encoding, expanded the dataset to 100 features. This processed dataset was then divided into training (785 rows) and test (337 rows) sets. These refined datasets were utilized for subsequent model training and evaluation.

3.5.1. *Model development & validation*

Table 7 presents the top five best-performing regressors for pchem and experimental datasets, evaluated using a 10-fold cross-validation process. For the pchem dataset, the Extra Trees regressor demonstrated superior performance, achieving the highest R^2 score (0.906), the lowest MSE (14.06), and lowest MAE (2.03). For the experimental dataset, the XGB regressor outperformed other models with an R^2 of 0.788, an MSE of 21.56, and an MAE of 2.86. While models such as CatBoost and RF regressors performed competitively, they did not surpass the performance of Extra Trees for pchem and XGB for the experimental dataset.

Fig. 9 shows the scatter plots depicting the performance of the top-performing models for the pchem dataset (left) and experimental dataset (right), with actual values plotted against predicted values. The Extra Trees Regressor for the pchem dataset shows strong predictive performance, as most points closely align with the diagonal, with minor underprediction at higher actual values. The XGB Regressor for the experimental dataset demonstrates reasonable predictive power but with greater variability around the diagonal, indicating higher prediction errors, particularly for larger values. Q-Q plots, residuals plots, and residual distributions of the models are shown in ***Fig. S5***.

3.5.2. *Feature importance*

Fig. 10 shows the feature importance charts illustrating the top 10 most influential features for best-performing models: (A) pchem dataset (Extra Trees Regressor) and (B) experimental dataset (XGB Regressor). Importance scores were derived using each model's algorithms. For Extra Trees importance is determined based on the Gini impurity reduction, calculated as the average decrease in impurity when a feature is used for splitting across all decision trees. For XGB Regressor importance is based on the frequency of feature usage in splits and the associated gain in model performance. For pchem dataset (Fig. 10, ***A)***, the exposure concentration ranks as the most important feature in predicting DNA damage. The cell culture medium and the cell line also significantly contribute, highlighting the role of the biological

Table 6
Top 10 feature importance of the entire dataset as determined by the Extra Trees Regressor (Gini importance). Higher-ranked features contribute more significantly to model predictions i.e., the relative contribution of each feature to the model's predictions performances.

Rank	Feature name	Importance score
1	Exposure Concentration (μg/mL)	0.198915
2	Cell Culture Medium (i.e., DMEM, RPMI 1640, BEGM, Hams F12, etc.,)	0.146866
3	Nanomaterial Type (i.e., NM100, Aeroxide P25, NM103, etc.,)	0.097088
4	Cell Line Name (A549, BEAS_2B, Caco2, TK6, others, etc.,)	0.065003
5	Positive Control used (H_2O_2, MMS, Ro 19–8022, EMS, etc)	0.059142
6	Cell Type (epithelial, hepatocytes, B-lymphoblasts, fibroblast, etc)	0.050663
7	Assay Type (alkaline comet, Fpg-modifed, net Fpg-modifed, etc)	0.046206
8	Electrophoresis Current (mA)	0.038568
9	Exposure Duration (h)	0.028804
10	Synthesis Method / Source (JRC, Sigma Aldrich, Evonik Degussa etc)	0.026587

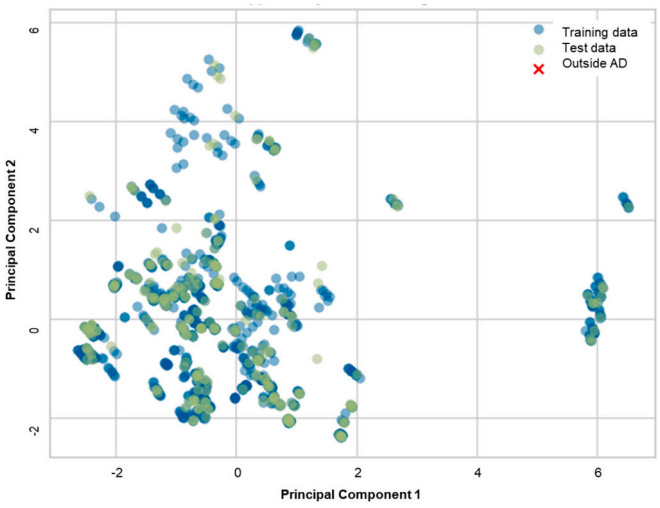


Fig. 8. Applicability Domain of the model as visualized using the PCA plot for the Extra Trees Regressor. The plot illustrates the distribution of training and test data points within the principal component space.

electrophoresis buffer (label) rank next, reflecting the role of buffer composition in the experimental design. Features such as NMs, staining compound, and Statistical approach chosen for the results analysis in the studies contribute to the model's predictive power, while the type of multiwell and assay type have lower importance, capturing minor variations in experimental setup.

3.5.3. Applicability domain

The applicability domains of the models are visualized in ***Fig. S6***. For the pchem dataset (***Fig. S6, left)***, most training and test data points are densely clustered near the center of the PCA space, indicating a well-defined applicability domain. Only a small number of test points fall outside this domain, suggesting these instances are less well-represented by the training data. The overall distribution is balanced, with minimal outliers, reflecting a relatively homogeneous data structure. In contrast, the experimental dataset (***Fig. S6, right)*** exhibits a more scattered distribution. This suggests a broader variability in the data, with a more dispersed applicability domain compared to the pchem dataset. While both datasets have a few points outside the applicability domain, the tighter clustering in the pchem dataset highlights its more uniform data structure, whereas the experimental dataset reflects greater heterogeneity and variability in feature space.

4. Discussion

Although no specific OECD guideline currently exists for the in vitro Comet assay, the relevance of this assay extends to regulatory frameworks, since it has been widely used to assess the genotoxic potential of various NMs. For example, it has been applied within the OECD Working Party on Manufactured Nanomaterials (WPMN) Testing Programme to evaluate TiO_2, silicon dioxide (SiO_2) NMs, and single-walled carbon nanotubes (SWCNTs). Moreover, the EU-funded RISKGONE project has proposed a standard operating procedure (SOP) for the in vitro comet assay, paving the way for its potential inclusion in an OECD TG[5] (El Yamani et al., 2022). Additionally, the assay is recognized as a suitable test under the Registration, Evaluation, Authorisation, and Restriction of Chemicals (REACH) regulation, is accepted by the EFSA, and is employed in screening novel pharmaceuticals and cosmetic ingredients (Azqueta and Dusinska, 2015; Vandghanooni and Eskandani, 2011). Despite its broad application, the use of genotoxicity data for regulatory assessment of NMs remains inconsistent, largely due to unclear data requirements and challenges in standardization (Doak et al., 2023b). A key regulatory consideration is the necessity to demonstrate cell exposure, as cellular uptake and intracellular distribution are crucial for appropriately interpreting negative genotoxicity results. Moreover, NM

Table 7
Top 5 best-performing regressors for the Pchem and Experimental datasets, evaluated using a 10-fold cross-validation process. The table summarizes the performance metrics, including R^2, Mean Squared Error (MSE), and Mean Absolute Error (MAE), highlighting the models' predictive accuracy for each dataset.

Pchem dataset	R^2	MSE	MAE	Experimental dataset	R^2	MSE	MAE
CatBoost Regressor	0.883	17.65	2.59	CatBoost Regressor	0.770	23.46	2.91
Random Forest Regressor	0.841	23.92	2.70	Gradient Boosting Regressor	0.759	24.57	3.47
Extra Trees Regressor	**0.906**	**14.06**	**2.03**	Random Forest Regressor	0.753	25.15	2.63
XGB Regressor	0.848	22.90	2.97	**XGB Regressor**	**0.788**	**21.56**	**2.86**
LGBM Regressor	0.834	25.02	2.98	Extra Trees Regressor	0.729	27.65	2.47

environment and cell type in the model's decision-making process. Features like assay type, cell type, and exposure duration show moderate importance, emphasizing the experimental context, while sonication medium and glutamine concentration in the cell culture have smaller but notable influence in the model's performance. For the experimental dataset (Fig. 10, ***B)***, the cold lysis buffer (label) is the most influential feature, underscoring the importance of lysis conditions in DNA damage prediction. The positive control used in the experiments and interference with assay components remains a recognized issue. While short-term exposure scenarios may not always detect systemic NM bioavailability, repeated exposures may lead to bioaccumulation in key organs, further complicating hazard assessment. There is, therefore, an

[5] https://www.enaloscloud.novamechanics.com/riskgone/material/material-TGs.html

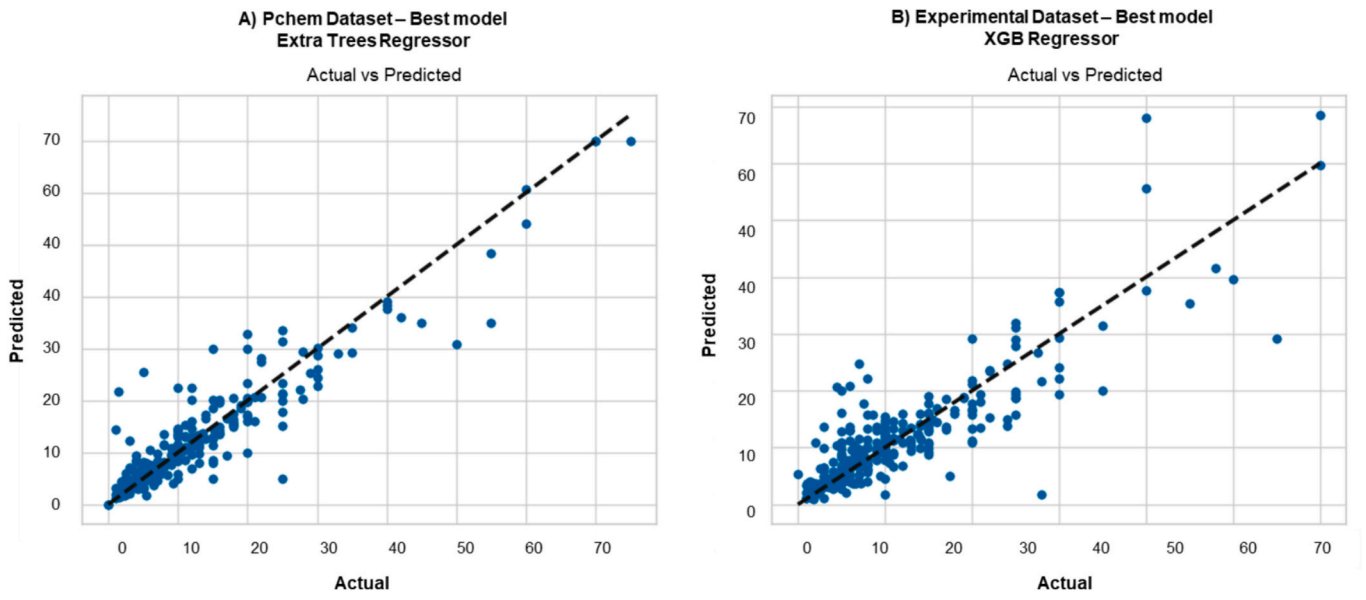


Fig. 9. Models' performance (actual vs predicted values) for the top-performing regressors. The left panel shows the scatter plot for the pchem dataset using the Extra Trees Regressor, and the right panel shows the scatter plot for the experimental dataset using the XGB Regressor.

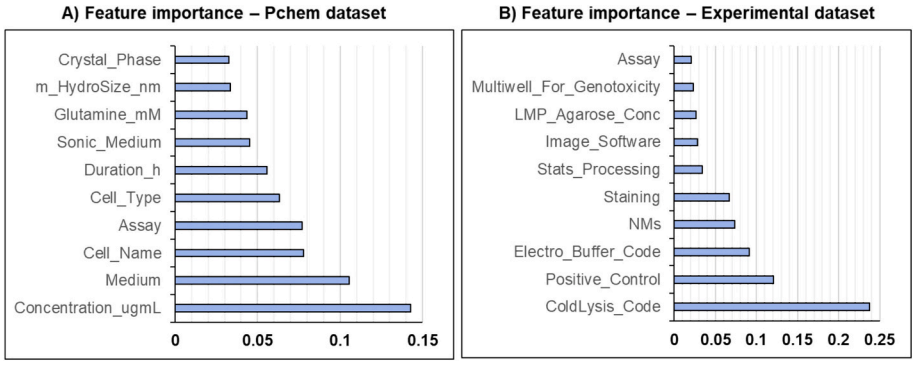


Fig. 10. Feature importance of the top 10 features derived from the best-performing models: Extra Trees Regressor for the pchem dataset (A) and XGB Regressor for the experimental dataset (B). These charts provide insight into the factors driving model predictions for each dataset.

urgent need for standardized approaches to NM pchem characterization, cellular interactions, and exposure validation to enhance regulatory confidence in genotoxicity data (Doak et al., 2023b; Allan et al., 2021).

Grouping and read-across strategies have been increasingly adopted within regulatory risk assessment, particularly under REACH, to enable the assessment of similar substances based on their shared properties. For NMs, read-across can be applied by integrating pchem properties, exposure metrics, and hazard endpoints to predict genotoxic potential across NMs (nanoforms) of the same or similar materials (Landsiedel et al., 2022b). Importantly, our study's modeling approach aligns with these regulatory principles, offering an *in-silico* NAM tool for NM hazard assessment through read-across methodologies. By applying ML to utilize genotoxicity data from multiple studies, our model provides a structured and reproducible framework to predict DNA damage potential based on pchem and experimental parameters. The ability of our dataset and models to capture key predictors of genotoxicity, such as concentration, exposure medium, and assay conditions, strengthens their utility in regulatory contexts. While our model serves as a proof-of-concept for regulatory applications, future enhancements such as incorporating cellular uptake metrics, NM interferences, and bioavailability data could further bridge the gap between computational hazard assessment and regulatory decision-making. Ultimately, this work represents a stepping stone toward integrating ML driven read-across approaches into regulatory nanosafety frameworks, providing a data-driven pathway for improving hazard assessment methodologies for NMs.

4.1. Modeling and datasets

This study assessed the performance of multiple ML models, including Extra Trees, XGB, Gradient Boosting and RF, across three datasets. The Extra Trees consistently outperformed other models for the entire and pchem datasets, achieving the highest R^2 values and the lowest MSE and MAE values. For the entire dataset, the Extra Trees model achieved an R^2 value of 0.84, reflecting a strong ability to explain the variance in DNA damage outcomes across diverse experimental and material conditions. Despite this strong performance, the model did not surpass the performance achieved with the pchem dataset (R^2 value of 0.90). In contrast, the experimental dataset, which isolated variables related to experimental design (e.g., imaging, and electrophoresis conditions), was best modelled by XGB. This model achieved an R^2 of 0.78, indicating predictive power, though lower than the pchem dataset. The relatively lower R^2 suggests that while experimental conditions enable the prediction of genotoxicity outcomes, they may not contribute as significantly as pchem properties in the predictions. These findings are in accordance also to previous work of Møller and colleagues (Møller

et al., 2015), who reported particle size, composition and crystal structure of NMs as important determinants of toxicity. This underscores the importance of prioritizing detailed pchem characterization, combined with more nuanced exposure parameters, in future studies. Such an approach is essential for improving the reliability of predictive models. Tree-based methods, such as Extra Trees and XGB, are recognized for their ability to handle non-linear relationships and capture intricate feature interactions (Ke et al., 2017). These models work by recursively partitioning the feature space into smaller, more homogeneous regions using decision trees. The Extra Trees creates an ensemble of decision trees by randomly selecting both features and data points, then averages the results across all trees. This randomization reduces variance, minimizes overfitting, and enhances the model's performance. The XGB Regressor is an implementation of gradient-boosted decision trees, optimized for accuracy. It builds models sequentially, with each tree correcting the errors of its predecessors, thereby improving performance. A key advantage of tree-based methods is their ability to automatically assess feature importance. This capability provides valuable insights into which variables have the greatest influence on models' outcomes, enabling researchers to prioritize key factors for further experimental analysis.

4.2. Feature importance

One of the challenges in predictive modeling is identifying the relevant features that contribute to accurate predictions. In in vitro nanotoxicology systems, numerous variables can be measured, but not all provide the same predictive power. Feature importance analysis serves as a fundamental step in model development, allowing to distinguish between influential and non-influential variables, thereby optimizing model performance and interpretability. In this study, our feature importance analysis revealed that concentration, exposure medium, and NM type were the most influential factors in predicting DNA damage outcomes. This finding aligns with prior research indicating that medium composition and experimental handling can significantly affect NM toxicity measurements (Dutta et al., 2007; Akabori and Nagle, 2014; Cullen et al., 2011; Pele et al., 2015; Pathakoti et al., 2013; Li et al., 2013; Ju et al., 2013). The type of TiO_2 also emerged as a key predictor, suggesting variability in genotoxicity across different NMs. Future modeling efforts could focus on specific NMs or tailor models to particular suspension media to improve predictive accuracy. Other important variables included cell type, positive controls, and electrophoresis conditions, suggesting that both pchem properties and experimental parameters influence genotoxicity responses. This is in accordance with the work of Carriere et al., in which the authors reported that different cell lines, alternative to the specific cells recommended by the expert community in genotoxicity (i.e. CHO-K1, V79 and TK6), could have a different sensitivity to genotoxicity that should be documented, alongside specific positive and historical controls (Carriere et al., 2020b). Moreover, more specific target organs representative also of the exposure routes toward NMs and more realistic exposure scenarios (air liquid interface exposure instead of classic submerged exposure) should also be considered for future genotoxicity assessments.

For the pchem dataset, features such as exposure duration and sonication medium were influential, suggesting that NM preparation and handling before exposure impact the genotoxic predictivity. Mu, Wang (Mu et al., 2019; Czajka et al., 2015), showed that the potential toxic effect of TiO_2-NMs can be related to duration of exposure. Hydrodynamic size and crystal phase of the NMs played smaller yet important roles (Murugadoss et al., 2020). The literature is inconsistent regarding the genotoxic potential of crystalline forms. For instance, studies like, Bernardeschi, Guidi (Bernardeschi et al., 2021) and Yu, Wang (Yu et al., 2017) report that pure anatase TiO_2-NMs are free of cyto- and genotoxicity, whereas Gao, Zhou (Gao et al., 2018), De Matteis, Cascione (De Matteis et al., 2016) and Chen, Yan (Chen et al., 2014) confirmed that anatase TiO_2-NMs show more toxicity than rutile forms.

In the experimental dataset, cold lysis code was identified as the most important feature. This highlights the role of the cold lysis step in the comet assay, as variations in lysis conditions, can influence the predicted damaged DNA (Tice et al., 2000). Other technical factors, such as the electrophoresis buffer and electrophoresis current, were also important, demonstrating that variations in the conditions under which DNA migration occurs can impact the predictability of DNA damage. Additional experimental parameters, including the LMP agarose concentration, and image software, played lesser but still notable roles in influencing model accuracy.

Across all datasets, concentration consistently stood out as the most influential feature, reinforcing its central role. However, in vitro dose is a complex issue as dose added to medium is different from delivered dose depending on media composition, volume, and NM concentration. In many cases mass precipitation occurs. Future research should incorporate a more refined representation of exposure dose as a feature. The kinetics of NMs in in vitro studies is often overlooked, despite their significant impact on dosimetry and hazard ranking. For instance, some NMs exhibit buoyancy due to low effective densities of their agglomerates in culture media, which affects particle transport, deposition and dose–response relationships. This can lead to underestimations of toxicity and bioactivity (Watson et al., 2016). The medium and cell line/type features were also highly influential in both the entire and pchem datasets, highlighting the importance of both the environment and biological variability in the genotoxicity of TiO_2-NMs (Brandão et al., 2020b).

4.3. Imputation strategy and data quality

Throughout this study, varying degrees of missing data were encountered across features. A thorough literature review revealed frequent underreporting of parameters, such as DNA unwinding temperature (67.5 % missing data), PdI, and zeta potential measurements in both water and cell medium (>50 % missing data). These features were subsequently excluded due to their substantial gaps. Similarly, sonication-related features (e.g., power, protocol, temperature, amplitude, and frequency) were often omitted in the studies, as were parameters related to light conditions, electrophoresis temperature, and cell density, necessitating their removal from the dataset. This underscores the need for improved reporting standards in future studies. Comprehensive documentation of sonication protocols, pchem characterization, and electrophoresis conditions should be prioritized, as these factors are essential for understanding NM dispersion and experimental reproducibility. Additionally, consistent reporting of cell density-related features, such as density per well or per cm^2, is essential to ensure replicability and comparability across studies. Despite these challenges, missing values were effectively handled using the MICE method. MICE allowed us to impute missing values by leveraging the relationships between variables, preserving the dataset's inherent structure and ensuring robustness (Mohammed et al., 2021). This approach outperforms simpler imputation methods. Given the proportion of missing data in pchem and experimental variables, imputing these values was necessary to preserve dataset usability. Although some parameters such as hydrodynamic size diameter or zeta potential can be reasonably inferred based on pchem trends, others such as sonication time or electrophoresis voltage are highly dependent on study-specific protocols. While this approach maximizes the dataset's utility, it does not replace the need for improved data reporting standards in nanosafety research. Future work should prioritize direct experimental validation and encourage more comprehensive metadata collection to minimize the need for imputation. Ensuring the completeness and reliability of pchem and hazard datasets is essential for improving the robustness of ML predictions in nanosafety assessments. Future work could and should also integrate data quality evaluation methodologies to assess dataset completeness and reliability, as highlighted in previous studies such as (Basei et al., 2022; Bossa et al., 2021b). Approaches like automated

completeness scoring systems and interactive feedback mechanisms, as proposed in the eNanoMapper database, could be valuable tools for enhancing dataset quality before modeling.

4.4. Scientific contributions

This study offers a FAIR (Findable, Accessible, Interoperable, Reusable) dataset, providing a valuable resource for advancing research and development in nanosafety (Jeliazkova et al., 2021; Papadiamantis et al., 2020). By adopting a data-driven approach to genotoxicity evaluation, this work demonstrates the potential of ML models in hazard assessment, supporting the broader integration of computational tools within nanosafety frameworks. The dataset aligns with current nanoinformatics efforts by offering a well-structured resource that facilitates the deployment of computational models (Serrano et al., 2024). Researchers can utilize this dataset to complement existing data, introduce novel features, or integrate additional studies to address data gaps and improve data quality. Its open accessibility encourages collaboration and data-sharing within the scientific community. Moving forward, expanding, and refining this dataset by addressing missing data and improving experimental consistency and data quality, will be essential for advancing the reliability of in silico nanosafety assessments.

4.5. Patents vs open science

An interesting parallel to our study lies in a recently submitted patent,[6] which developed a QSAR model (Method B) for predicting the genotoxicity of TiO_2 nanoforms using a classification approach. Their method relies on pchem descriptors and restricts access to data and findings through patent protection. In contrast, our approach employs a regression-based model, providing quantitative predictions of DNA damage percentages. By incorporating a significantly larger set of descriptors, our model captures a broader spectrum of pchem properties and experimental conditions, thereby enhancing its generalizability. Unlike the proprietary nature of the patented QSAR model, our study aligns with the principles of open science by sharing a FAIR dataset (the python code is available upon request to the author). This data openness facilitates collaboration and enables other researchers to extend our work, refine the models, and contribute to the development of safer NMs. This comparison underscores the value of our approach and also its ethos of open science and collaboration, which we believe will accelerate progress in the field of nanoinformatics.

4.6. Considerations and future directions

In vivo studies were excluded from this study because a harmonized exposure metric system is needed to transpose in vivo exposure dose (e.g., milligrams per kilogram of body weight) to a quantity compatible to the target tissue dose (biologically effective dose) of in vitro studies. In addition, the experimental settings would differ across the studies, for example for the in vivo studies the duration of exposure, the route of administration, (a substance have different effects when it is administered orally versus intravenously) and the use of other substances have a significant impact on the results and are not directly comparable with the in vitro features. Recent studies have demonstrated that in vitro toxicity tests can provide credible results, making them a valuable tool for toxicity testing approaches, offering several benefits, including greater control over experimental conditions and less ethical concerns associated with animal use (Burden et al., 2021). We selected in vitro approaches since they are more abundant in the literature in comparison to in vivo studies and also encourage the 3Rs principles (Replacement, Reduction, and Refinement). While in vitro studies have been instrumental in elucidating the genotoxicity of TiO_2-NMs, a comprehensive understanding of the underlying mechanisms necessitates in vivo investigations. Nevertheless, the translation of in vitro results to an in vivo environment warrants careful consideration due to certain intricacies. For example, after in vitro exposure to TiO_2 NM, in the cell cytoplasm could be found large agglomerates that could interfere with the migration of DNA during electrophoresis. This example points out that more key features should be considered (i.e. uptake of NMs) (Carriere et al., 2020a). Moreover, in the past the use of comet assay was performed without automatic scoring. Results could have been also influenced by single operators scoring methods, leading to misinterpretation of the results. Although its limitations, the comet test has been used in several application and papers to assess to genotoxicity of NMs, not only in vitro for human hazard assessment but also for environmental hazard assessment purposes.

While no OECD guideline exists for the in vitro comet assay, its widespread application aligns with efforts to promote non-animal testing methods (Pfuhler et al., 2020). A compendium of protocols provided by Collins, Møller (Collins et al., 2023) outlines the assay's advantages, challenges, and limitations in genotoxicity assessment, while Møller, Hemmingsen (Møller et al., 2015) highlight the methodological challenges encountered in particle toxicology. The comet assay, including enzyme-modified versions, has proven valuable for detecting both DNA strand breaks and oxidative damage (Collins et al., 2023; García-Rodríguez et al., 2019). These enzyme-modified variants enhance the sensitivity of the assay, enabling a more comprehensive evaluation of DNA damage. Such advancements contribute to the understanding of key events within Adverse Outcome Pathways (AOPs), providing mechanistic insights into the progression from molecular-level damage to adverse effects, such as lung carcinogenicity, particularly relevant in NM exposure. Our work supports the broader recognition of the in vitro comet assay as a key component within integrated approaches to testing and assessment (IATA, structured frameworks that combine various sources of information such as in vitro assays, in silico models, and existing data to support regulatory decision-making) (Bossa et al., 2021a). The integration of in silico NAMs, such as ML models, can complement traditional in vitro assays by offering predictive capabilities. These models can support IATA frameworks by identifying key factors that influence genotoxicity, aiding in the prioritization of experimental testing and facilitating read-across approaches for regulatory assessments.

5. Conclusion

This study highlights the role of ML models in nanosafety by predicting the genotoxic potential of TiO_2 NMs. By systematically compiling a dataset from in vitro comet assay studies, integrating pchem properties, exposure conditions, and experimental design variables, we developed predictive read across models that achieve high accuracy. The successful application of tree-based models, particularly Extra Trees Regressor and XGB Regressor, demonstrated their ability to capture nonlinear relationships and complex interactions between experimental and pchem features. This study provides insights into the key factors influencing the prediction of the genotoxicity of TiO_2 NMs. Concentration, the cell suspension medium, cell type, and NM type were found to be influential. The study also emphasizes the importance of experimental procedures, especially in relation to key variables like cold lysis and electrophoresis, which can significantly impact the predictions of genotoxicity outcomes. The study also stressed the need for improved reporting results in the literature. Approximately 50 % of the columns were removed due to high proportions of missing data (>30 %), including key features like NMs' PdI, zeta potential, sonication parameters, and cell density-related measurements. This work contributes to the growing body of literature promoting non-animal testing methods. Moreover, the development of a FAIR dataset enhances transparency,

[6] https://www.mysciencework.com/patent/show/methods-prediction-titanium-dioxide-nanoparticles-mutagenic-genotoxic-effects-human-health-EP4362025A1

encourages collaboration, and facilitates future advancements in nanosafety research.

CRediT authorship contribution statement

Irini Furxhi: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Mahsa Mirzaei:** Writing – review & editing, Writing – original draft, Investigation, Data curation, Conceptualization. **Anna Costa:** Writing – review & editing, Funding acquisition. **Rossella Bengalli:** Writing – review & editing, Writing – original draft.

Funding

This research was funded by the European Union's Horizon MSCA-2022-PF-01-01 Programme, grant N°101103082.

Declaration of competing interest

The authors declare no conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.impact.2025.100562.

Data availability

Dataset available as supplementary material and in Zenodo: https://doi.org/10.5281/zenodo.15057424

References

Abend, A., et al., 2024. Industry's perspective on challenges assessing the in vivo impact of removing titanium dioxide (TiO2) from drug products. J. Pharm. Sci. 113 (11), 3119–3122.
Abubaker, H., et al., 2024. Exploring important factors in predicting heart disease based on ensemble- extra feature selection approach. Baghdad Sci. J. 21 (2(SI)), 0812.
Additives, E., Panel, O.F., et al., 2021. Safety assessment of titanium dioxide (E171) as a food additive, 19 (5), e06585.
Akabori, K., Nagle, J.F., 2014. Comparing lipid membranes in different environments. ACS Nano 8 (4), 3123–3127.
Allan, J., et al., 2021. Regulatory landscape of nanotechnology and nanoplastics from a global perspective. Regul. Toxicol. Pharmacol. 122, 104885.
Andreoli, C., et al., 2018. Critical issues in genotoxicity assessment of TiO(2) nanoparticles by human peripheral blood mononuclear cells. J. Appl. Toxicol. 38 (12), 1471–1482.
Armand, L., et al., 2016. Long-term exposure of A549 cells to titanium dioxide nanoparticles induces DNA damage and sensitizes cells towards genotoxic agents. Nanotoxicology 10 (7), 913–923.
Azqueta, A., Dusinska, M., 2015. The use of the comet assay for the evaluation of the genotoxicity of nanomaterials. Front. Genet. 6, 239.
Basei, G., et al., 2022. A methodology for the automatic evaluation of data quality and completeness of nanomaterials for risk assessment purposes. Nanotoxicology 16 (2), 195–216.
Bernardeschi, M., et al., 2021. Suitability of nanoparticles to face benzo(a)pyrene-induced genetic and chromosomal damage in M. Galloprovincialis. An In Vitro Approach. Nanomaterials 11 (5), 1309.
Bossa, C., et al., 2021a. FAIRification of nanosafety data to improve applicability of (Q) SAR approaches: a case study on in vitro comet assay genotoxicity data. Computat. Toxicol. (Amsterdam, Netherlands) 20, 100190.
Bossa, C., et al., 2021b. FAIRification of nanosafety data to improve applicability of (Q) SAR approaches: a case study on in vitro comet assay genotoxicity data. Computat. Toxicol. 20, 100190.
Brandão, F., et al., 2020a. Genotoxicity of TiO2 nanoparticles in four different human cell lines (A549, HEPG2, A172 and SH-SY5Y). Nanomaterials 10 (3), 412.
Brandão, F., et al., 2020b. Genotoxicity of TiO2 Nanoparticles in Four Different Human Cell Lines (A549, HEPG2, A172 and SH-SY5Y), 10 (3), 412.
Burden, N., et al., 2021. Opportunities and challenges for integrating new in vitro methodologies in Hazard testing and risk assessment. Small 17 (15), 2006298.
Carriere, M., Arnal, M.-E., Douki, T., 2020a. TiO2 genotoxicity: an update of the results published over the last six years. Mutat. Res./Genet. Toxicol. Environ. Mutagen. 854-855, 503198.
Carriere, M., Arnal, M.E., Douki, T., 2020b. TiO(2) genotoxicity: an update of the results published over the last six years. Mutat. Res. Genet. Toxicol. Environ. Mutagen. 854-855, 503198.
Charles, S., et al., 2018. Assessment of the in vitro genotoxicity of TiO2 nanoparticles in a regulatory context. Nanotoxicology 12 (4), 357–374.
Chen, T., Guestrin, C., 2025. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. Association for Computing Machinery, San Francisco, California, USA, pp. 785–794.
Chen, T., Yan, J., Li, Y., 2014. Genotoxicity of titanium dioxide nanoparticles. J. Food Drug Anal. 22 (1), 95–104.
Chen, Z., et al., 2022. DNA oxidative damage as a sensitive genetic endpoint to detect the genotoxicity induced by titanium dioxide nanoparticles. Nanomaterials 12 (15), 2616.
Chicco, D., Warrens, M.J., Jurman, G., 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput. Sci. 7, e623.
Collins, A., et al., 2023. Measuring DNA modifications with the comet assay: a compendium of protocols. Nat. Protoc. 18 (3), 929–989.
Commission, E, et al., 2019. An overview of concepts and terms used in the European Commission's definition of nanomaterial. Publications Office.
Cort, J.W., Kenji, M., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 30 (1), 79–82.
Cullen, L.G., et al., 2011. Assessing the impact of nano-and micro-scale zerovalent iron particles on soil microbial activities: particle reactivity interferes with assay conditions and interpretation of genuine microbial effects. Chemosphere 82 (11), 1675–1682.
Czajka, M., et al., 2015. Toxicity of titanium dioxide nanoparticles in central nervous system. Toxicol. in Vitro 29 (5), 1042–1052.
De Matteis, V., et al., 2016. Toxicity assessment of anatase and rutile titanium dioxide nanoparticles: the role of degradation in different pH conditions and light exposure. Toxicol. in Vitro 37, 201–210.
de Myttenaere, A., et al., 2016. Mean absolute percentage error for regression models. Neurocomputing 192, 38–48.
Doak, S.H., et al., 2012. In vitro genotoxicity testing strategy for nanomaterials and the adaptation of current OECD guidelines. Mutat. Res. 745 (1–2), 104–111.
Doak, S.H., et al., 2023a. Current status and future challenges of genotoxicity OECD test guidelines for nanomaterials: a workshop report. Mutagenesis 38 (4), 183–191.
Doak, S.H., et al., 2023b. Current status and future challenges of genotoxicity OECD test guidelines for nanomaterials: a workshop report. Mutagenesis 38 (4), 183–191.
Dutta, D., et al., 2007. Adsorbed proteins influence the biological activity and molecular targeting of nanomaterials. Toxicol. Sci. 100 (1), 303–315.
El Yamani, N., et al., 2022. The miniaturized enzyme-modified comet assay for genotoxicity testing of nanomaterials, 4.
Elje, E., et al., 2020. Hepato(Geno)Toxicity Assessment of Nanoparticles in a HepG2 Liver Spheroid Model, 10 (3), 545.
Fatima, S., Yadam, S., 2023. The Comet Assay: A Straight Way to Estimate Geno-Toxicity, vol. 3(2). 21st Century Pathology, p. 145.
Furxhi, I., et al., 2019. Machine learning prediction of nanoparticle in vitro toxicity: a comparative study of classifiers and ensemble-classifiers using the Copeland index. Toxicol. Lett. 312, 157–166.
Furxhi, I., et al., 2020a. Practices and trends of machine learning application in nanotoxicology. Nanomaterials 10 (1), 116.
Furxhi, I., et al., 2020b. Nanotoxicology data for in silico tools: a literature review. Nanotoxicology 14 (5), 612–637.
Gao, X., et al., 2018. Distinct effects of soluble and bound exopolymeric substances on algal bioaccumulation and toxicity of anatase and rutile TiO 2 nanoparticles. Environ. Sci. Nano 5 (3), 720–729.
García-Rodríguez, A., et al., 2019. The comet assay as a tool to detect the genotoxic potential of nanomaterials, 9 (10), 1385.
Gázquez, M.J., et al., 2014. A review of the production cycle of titanium dioxide pigment. Mater. Sci. Appl. 2014, 441–458.
Gázquez, M.J., Moreno, S.M.P., Bolívar, J.P., 2021. TiO2 as white pigment and valorization of the waste coming from its production. In: Titanium Dioxide (Tio_2) and its Applications. Elsevier, pp. 311–335.
Ghosh, M., Chakraborty, A., Mukherjee, A., 2013. Cytotoxic, genotoxic and the hemolytic effect of titanium dioxide (TiO2) nanoparticles on human erythrocyte and lymphocyte cells in vitro, 33 (10), 1097–1110.
Jeliazkova, N., et al., 2021. Towards FAIR nanosafety data. Nat. Nanotechnol. 16 (6), 644–654.
Ju, L., et al., 2013. Quantum dot-related genotoxicity perturbation can be attenuated by PEG encapsulation. Mutat. Res./Genet. Toxicol. Environ. Mutagen. 753 (1), 54–64.
Ke, G., et al., 2017. LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Long Beach, California, USA, pp. 3149–3157.
Khan, S.I., Hoque, A.S.M.L., 2020. SICE: an improved missing data imputation technique. J. Big Data 7 (1), 37.
Kim, Y., Byun, Y.-C., Lee, S.-J., 2024. A Study on Sugar Content Improvement and Distribution Flow Response through Citrus Sugar Content Prediction Based on the PyCaret Library, 10 (6), 630.
Kirkland, D., et al., 2022. A weight of evidence review of the genotoxicity of titanium dioxide (TiO(2)). Regul. Toxicol. Pharmacol. 136, 105263.
Kurita, T., 2019. Principal component analysis (PCA). In: Computer Vision: A Reference Guide. Springer International Publishing, Cham, pp. 1–4.

Lamon, L., et al., 2018. Grouping of nanomaterials to read-across hazard endpoints: from data collection to assessment of the grouping hypothesis by application of chemoinformatic techniques. Part. Fibre Toxicol. 15 (1), 37.
Landsiedel, R., et al., 2022a. Genotoxicity testing of nanomaterials. Wiley Interdiscip. Rev. Nanomed. Nanobiotechnol. 14 (6), e1833.
Landsiedel, R., et al., 2022b. Genotoxicity testing of nanomaterials, 14 (6), e1833.
Li, R., et al., 2013. Surface charge and cellular processing of covalently functionalized multiwall carbon nanotubes determine pulmonary toxicity. ACS Nano 7 (3), 2352–2368.
Liang, C., et al., 2024. Genotoxicity evaluation of food additive titanium dioxide using a battery of standard in vivo tests. Regul. Toxicol. Pharmacol. 148, 105586.
Lu, Y., Liu, Y., Yang, C., 2017. Evaluating in vitro DNA damage using comet assay. J. Vis. Exp. 128.
Mera-Gaona, M., et al., 2021. Evaluating the impact of multivariate imputation by MICE in feature selection. PLoS One 16 (7), e0254720.
Mohammed, M.B., et al., 2021. Comparison of five imputation methods in handling missing data in a continuous frequency table. AIP Conf. Proc. 2355(1).
Møller, P., et al., 2015. Applications of the comet assay in particle toxicology: air pollution and engineered nanomaterials exposure. Mutagenesis 30 (1), 67–83.
Møller, P., et al., 2020. Minimum information for reporting on the comet assay (MIRCA): recommendations for describing comet assay procedures and results. Nat. Protoc. 15 (12), 3817–3826.
Mu, W., et al., 2019. Effect of long-term intake of dietary titanium dioxide nanoparticles on intestine inflammation in mice. J. Agric. Food Chem. 67 (33), 9382–9389.
Murugadoss, S., et al., 2020. Agglomeration of titanium dioxide nanoparticles increases toxicological responses in vitro and in vivo. Part. Fibre Toxicol. 17 (1), 10.
Muruzabal, D., Collins, A., Azqueta, A., 2021. The enzyme-modified comet assay: past, present and future. Food Chem. Toxicol. 147, 111865.
OECD, 2016. Test No. 490. In: Vitro Mammalian Cell Gene Mutation Tests Using the Thymidine Kinase Gene.
OECD, 2016a. Test No. 476: In Vitro Mammalian Cell Gene Mutation Tests using the Hprt and xprt genes.
OECD, 2016b. Test No. 473: In Vitro Mammalian Chromosomal Aberration Test.
OECD, 2023. Test No. 487: In Vitro Mammalian Cell Micronucleus Test.
Papadiamantis, A.G., et al., 2020. Metadata stewardship in nanosafety research: community-driven organisation of metadata schemas to support FAIR nanoscience data, 10 (10), 2033.
Pathakoti, K., et al., 2013. In vitro cytotoxicity of CdSe/ZnS quantum dots with different surface coatings to human keratinocytes HaCaT cells. J. Environ. Sci. 25 (1), 163–171.
Pele, L., et al., 2015. Artefactual nanoparticle activation of the inflammasome platform: in vitro evidence with a nano-formed calcium phosphate. Nanomedicine 10 (9), 1379–1390.
Pfuhler, S., et al., 2020. Validation of the 3D reconstructed human skin comet assay, an animal-free alternative for following-up positive results from standard in vitro genotoxicity assays. Mutagenesis 36 (1), 19–35.
Reddy, G.T., et al., 2020. Analysis of dimensionality reduction techniques on big data. Ieee Access 8, 54776–54788.
Regonia, P.R., et al., 2022. Machine learning-enabled nanosafety assessment of multi-metallic alloy nanoparticles modified TiO2 system. NanoImpact 28, 100442.
Ren, J., et al., 2022. Balanced MSE for Imbalanced Visual Regression.
Sajid, M., et al., 2015. Impact of nanoparticles on human and environment: review of toxicity factors, exposures, control strategies, and future prospects. Environ. Sci. Pollut. Res. Int. 22 (6), 4122–4143.
Sang, L., et al., 2022. Machine learning for evaluating the cytotoxicity of mixtures of Nano-TiO2 and heavy metals: QSAR model apply random Forest algorithm after clustering analysis. Molecules 27 (18), 6125.
Sayes, C.M., Reed, K.L., Warheit, D.B., 2007. Assessing toxicity of fine and nanoparticles: comparing in vitro measurements to in vivo pulmonary toxicity profiles. Toxicol. Sci. 97 (1), 163–180.
Schober, P., Boer, C., Schwarte, L.A., 2018. Correlation coefficients: appropriate use and interpretation. Anesth. Analg. 126 (5), 1763–1768.
Serrano, B.A., et al., 2024. The role of FAIR nanosafety data and nanoinformatics in achieving the UN sustainable development goals: the NanoCommons experience. RSC Sustainabilit. 2 (5), 1378–1399.
Sewell, F., et al., 2024. New approach methodologies (NAMs): identifying and overcoming hurdles to accelerated adoption. Toxicol. Res. (Camb) 13 (2) p. tfae044.
Singh, N.P., et al., 1988. A simple technique for quantitation of low levels of DNA damage in individual cells. Exp. Cell Res. 175 (1), 184–191.
Smiti, A., 2020. A critical overview of outlier detection methods. Comput Sci Rev 38, 100306.
Stone, V., Johnston, H., Schins, R.P., 2009. Development of in vitro systems for nanotoxicology: methodological considerations. Crit. Rev. Toxicol. 39 (7), 613–626.
Teasdale, A., Hughes, K., 2023. Regulatory highlights. Org. Process. Res. Dev. 27 (3), 394–398.
Tice, R.R., et al., 2000. Single cell gel/comet assay: Guidelines for in vitro and in vivo genetic toxicology testing, 35 (3), 206–221.
Trinh, T.X., et al., 2022. Developing random forest based QSAR models for predicting the mixture toxicity of TiO2 based nano-mixtures to Daphnia magna. NanoImpact 25, 100383.
Vandghanooni, S., Eskandani, M., 2011. Comet assay: a method to evaluate genotoxicity of nano-drug delivery system. Bioimpacts 1 (2), 87–97.
Wang, Z., Chen, J., 2023. Applicability domain characterization for machine learning QSAR models. In: Hong, H. (Ed.), Machine Learning and Deep Learning in Computational Toxicology. Springer International Publishing, Cham, pp. 323–353.
Watson, C.Y., et al., 2016. Buoyant Nanoparticles: Implications for Nano-Biointeractions in Cellular Studies, 12 (23), 3172–3180.
Weir, A., et al., 2012. Titanium dioxide nanoparticles in food and personal care products. Environ. Sci. Technol. 46 (4), 2242–2250.
Yu, Q., et al., 2017. Different toxicity of anatase and rutile TiO2 nanoparticles on macrophages: involvement of difference in affinity to proteins and phospholipids. J. Hazard. Mater. 335, 125–134.
Yu, L., et al., 2022. Missing data preprocessing in credit classification: one-hot encoding or imputation? Emerg. Mark. Financ. Trade 58 (2), 472–482.
Yuan, B., et al., 2021. QNAR modeling of cytotoxicity of mixing nano-TiO2 and heavy metals. Ecotoxicol. Environ. Saf. 208, 111634.
Ziental, D., et al., 2020. Titanium dioxide nanoparticles: prospects and applications in medicine. Nanomaterials 10 (2), 387.