Physical Education, Health and Social Sciences
https://journal-of-social-education.org
E-ISSN: 2958-5996
P-ISSN: 2958-5988

Machine Learning-Based Dual-Target Prediction of Metal Oxide Nanoparticle Cytotoxicity: Integrating Physicochemical, Electronic, and Compositional Descriptors with Transfer Learning

Hafsa Batool[1], Saeed Rasheed[2], Hamda Khalid[3], Samavia Khalid[4]

[1] Department of Physics, University of Agriculture Faisalabad, Pakistan
[2] Department of Computer Science, University of Agriculture Faisalabad, Pakistan
[3,4] Faculty of CS & IT The Superior University Lahore, Pakistan,
Email: hamdakhalid@superior.edu.pk, samaviakhalid@superior.edu.pk

DOI: https://doi.org/10.63163/jpehss.v4i2.1470

Abstract
There is an urgent need for robust and scalable computational tools for nanosafety assessment, given that the number of engineered metal oxide nanoparticles (NPs) is increasing rapidly in their use in industries and tissues across a wide variety of applications in biomedical sciences. Using the S2NANO MeOx_I meta-analysis, we consider 26 metal oxide NP materials for their in vitro cytotoxicity effect on both human and bacterial cells ($n = 6,842$ experimental records) and present an extensive machine learning framework to predict the dual target in vitro cytotoxicity. Our framework treats both binary toxicity classification (Toxic/Nontoxic) and continuous cell viability regression (Viability %) problems at the same time under strict out-of-distribution (OOD) evaluation protocol which includes a test set of novel materials not included in the training set. To systematically compare four baseline models (XGBoost and Random Forest (RF) for both tasks with two dual-target multilayer perceptron (MLP) architectures (A) physicochemical-feature MLP and (B) composition-embedding transfer learning MLP which uses 132-dimensional Magpie compositional embeddings derived from matminer library. Exploratory data analysis showed strong class imbalance (84.5% Nontoxic), an 8-order-of-magnitude range that required log-transformation, and monotonicity in the response appeared to be present. Surface formation enthalpy (H_{sf}) and valence band maximum (E_v) are among these key material electronic descriptors, which were consistently found to be among the most important in SHAP and Integrated Gradient analyses, in addition to $\log_{10}$(dose). In the case of the OOD test set, the best baseline classifier (RF) resulted in an ROC-AUC = 0.721 and Model B had an ROC-AUC = 0.737. To measure regression, Model B performed with $R^2 = 0.085$ with RMSE = 26.7% which is significantly better than all baselines (RF $R^2 = -0.060$ and RMSE = 28.8%) and utilized compositional transfer information. More importantly, model B achieved regression R^2 of +0.059 on Fe_3O_4 against all test materials, compared with −0.115 in the model RF. This illustrates that compositional embedding yield gains in both generalization and precision of the model, on chemically dissimilar test materials. The results made the creation of a reproducible NP nanosafety predictive baseline and showed the importance of electronic and compositional descriptors, which go beyond classical physicochemical descriptors.

Keywords: Nanotoxicity; Metal Oxide Nanoparticles; Machine Learning; Xgboost; Random Forest; Multilayer Perceptron; Transfer Learning; Compositional Embeddings; Shap; Integrated Gradients; Dual-Target Prediction; Out-Of-Distribution Generalization

Introduction
Engineered nanomaterials have been rapidly produced and applied all over the world over the last. Since a couple of decades, the metal oxide nanoparticles (NPs) have been in the middle of the list of the applications delivered from introduction to a few: Span from improving food packaging to

photovoltaic and photovoltaic energy storage and uses in rechargeable batteries. To drug delivery and medical imaging, photocatalysis and antimicrobial coatings [1,2]. Despite their, the cytotoxic effect of metal oxide NPs brings serious concerns to human health, particularly due to its technological promise. Along with this, ecological systems served as a motivator for the creation of environmentally friendly and cost-effective nanosafety assessment tools with high speed and efficiency capabilities [3,4]. Labor-intensive traditional vitro toxicology assays, which are essential, have drawbacks of inter Unable to scale to the combinatorial space of NP compositions - sizes, surface - in the laboratory and in practice, unable to scale to the number and the diversity of potential interfaces. In the laboratory and practice, unable to scale to numbers of NPs and to the diversity of interfaces. Real-world exposure conditions, which represent chemistry, and nanosized particle diversity [5]. Computational methods, and, more specifically, machine learning (ML), have proven to be valuable tools and complements. experimental nanotoxicology [6,7]. By learning structure-activity relationships from curated experimental datasets, ML models can predict toxicity endpoints of new NPs without going through extra assays, which also reduces the cost of the analysis. Datasets, ML models can predict toxicity endpoints of new NPs without going through extra assays, which also reduces toxicity analysis cost. Boosting nanosafety screening and regulatory risk assessment [8]. The development of reliable There are some longstanding basic problems in handling ML models for nanotoxicity: (i) High dimensional, and heterogeneous feature sets, (ii) Short experts timescale and most of the time spent on data cleaning, (iii) Minimal knowledge about each sample. approximation & goal BADNESS (i) impossible, (ii) severe class imbalance between the two classes. (iii) minimal numbers of datasets compared to the variety of nanmaterial chemistry; and moreover, (iv) the critical need for genuine out of distribution (OOD) generalisation; i.e. prediction over toxicity, is recognised. for materials do not present in the training set [9,10]. The S2NANO MeOx_I meta-analysis data set [11] which consists of 6,842 in vitro cytotoxicity datasets. Data on 26 metal oxide NP materials from the peer-reviewed literature offers an unprecedented resource for. To tackle these problems then in a scalable way. Its breadth of physicochemical descriptors (core size, hydrodynamic size, band maximum Ev), and charge transfer parameters (the number of electronic transitions). Charges, conductivity electron band minimum Ec, valence band maximum Ev, surface area, and quantum-chemical electronic characteristics (the number of electronic transitions). H_2, band maximum E_v, heat of surface formation H_s^f, metal-oxygen bond energy MeO), and experimental values of the atomic number Z. The metal-oxygen bond energy MeO) and experimental values of the atomic number Z. It is particularly well suited for multi-descriptor ML modelling thanks to the covariate information included (dose, exposure time, cell type, assay). Past computational investigations of the toxicity of metal oxide NPs are largely based on classical ML. microscale datasets of Three Graces mosasaurus small-scale mosasaurus data of a single material (Three Graces). Limited feature sets were used for the datasets [12,13]. Later, there was research on deep learning architecture, and the various approaches remain unexplored. Transfer learning approaches for predicting properties of nanomaterials [14,15] were examined, however, a comprehensive comparative study of the different approaches is lacking. However, there are few studies for baselines and transfer-learning models that adequately test the model against rigorous OOD. Furthermore, Ingestion study is commonly divided into two parts: prediction of binary toxicity classification and prediction of continuous viability regression—a couple. A common ingestion study involves the simultaneous prediction of both binary toxicity classification and continuous viability regression – a couple. Very little work has been done about the optimal formulation of targets to maximise information obtained from experimental data. Their research has been focused on the study of nanotoxicology, an area that has garnered attention in the nanotoxicology literature [16].

In this study, we address these gaps through a comprehensive, multi-phase ML investigation of the S2NANO MeOx_I dataset. Our contributions are as follows:

- A rigorous exploratory data analysis (EDA) pipeline encompassing missing value characterization, KNN imputation of measurement-method metadata, IQR-based outlier clipping, and material-level stratified train/test splitting that enforces strict OOD evaluation.

- Dual-target baseline models (XGBoost and Random Forest) for simultaneous toxicity classification and viability regression, evaluated using leave-one-material-out (LOMO) cross-validation and held-out OOD test materials.
- A dual-target MLP architecture (Model A) trained on physicochemical and electronic descriptors, and a transfer-learning MLP (Model B) that incorporates 132-dimensional Magpie compositional embeddings to encode material-level chemical information beyond hand-crafted features.
- Comprehensive feature attribution analyses using SHAP (for three models) and Captum Integrated Gradients (for neural networks), providing mechanistic insight into the physicochemical and electronic drivers of metal oxide NP cytotoxicity.
- A systematic comparison of all models on six OOD test materials, with particular focus on Fe_3O_4—the largest and most challenging test material—demonstrating the generalisation advantage of compositional transfer learning.

Methodology

Dataset and Source

We use the S2NANO MeOx_I meta-analysis dataset (DOI: 10.5281/zenodo.15300193), an all of them. Tables showing in vitro cytotoxicity data for Metal Oxide Engineered Nanomaterials that have been collated and curated. The raw There were 36 columns and 6,843 records initially in the dataset, one was completely removed because there were no usable records. null placeholder row. The data set covers 26 different types of metal oxide NP materials: Al_2O_3, Bi_2O_3, CeO_2, Co_2O_3, ... Co_3O_4, CoO, Cr_2O_3, CuO, Fe_2O_3, Fe_3O_4, Gd_2O_3, HfO_2, In_2O_3, La_2O_3, Mn_2O_3, Ni_2O_3, NiO, Sb_2O_3, SiO_2, SnO_2, TiO_2, V_2O_3, WO_3, Y_2O_3, Yb_2O_3, ZnO, and ZrO_2. The five most-represented materials are Fe_3O_4 ($n = 1,563$), TiO_2 ($n = 1,495$), ZnO ($n = 1,243$), SiO_2 ($n = 890$), and CeO_2 ($n = 156$). (> 50% cell viability), and (2) toxicity level, where toxicity of concentrations above 50% cell viability is ranked based on UV absorbance. toxic/nontoxic (≤ 50% cell viability vs. > 50% cell viability), and toxicity level based on absorbance of UV concentration above 50% cell viability. (1) cell viability decline for continued culture (3) cell viability (cell count/control) decline for continued culture (Viability %). The dataset has a high positive skew. class imbalance: 84.5% Nontoxic ($n = 5,783$) vs. 15.5% Toxic ($n = 1,059$), with mean viability 80.3% (median 90.6%, SD 28.6%). The distribution of the viability is well skewed to the left, and its shape is bell-shaped but with a very heavy tail towards the high end. High dose cytotoxic response in left tail.

Exploratory Data Analysis and Preprocessing

Missing Value Analysis and KNN Imputation

No values were missing in the other columns, except for four columns where the measurement method was explicitly indicated with respect to each physicochemical property, which showed the following missing values: Method_surface_area: 76.5%, Method_surface_charge: 43.3%, Method_hydro_size: 40.0%, and Method_core_size: 23.5%. The columns containing numeric physicochemical and electronic descriptors were completely filled with 100% completeness level. The missingness pattern was similar to that of Missing Completely At Random (MCAR), in that no systematic order of the presence of missing entries across materials and experimental conditions was found.

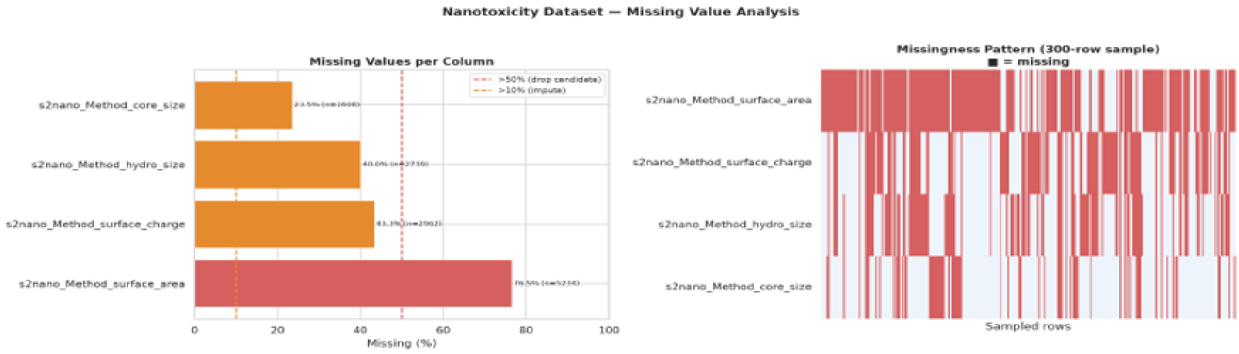


Figure 1: Missing Values

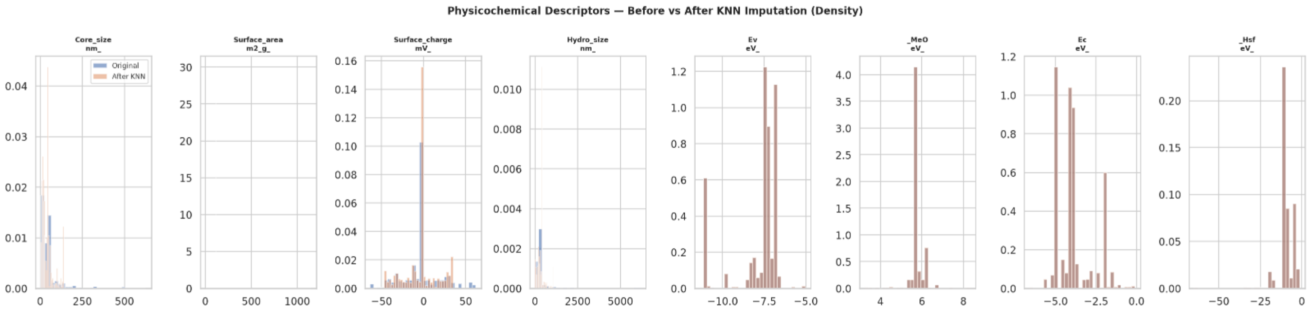


Figure 2: KNN Imputation

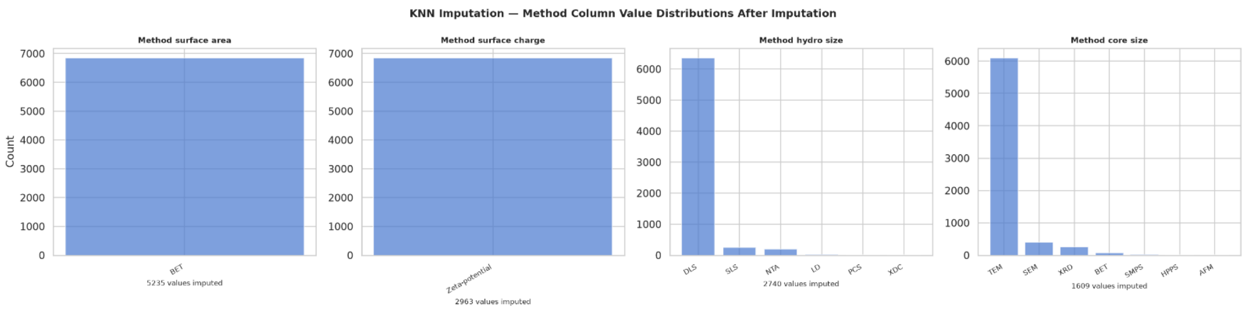


Figure 3: KNN Imputation on Methods

The four categorical method columns were label-encoded and then imputed using the KNN (k=5) imputation method. The integer codes for each class were rounded to the nearest integer code number, after which they were mapped back to the category for strings. There were a total of 12,543 values imputed in the four columns to provide a complete data set for downstream modelling.

Unit Standardization and Outlier Clipping

It was confirmed that all the physicochemical descriptors are in proper SI, i.e., dimension. Outliers were detected, using IQR-based clipping, which includes the computation of the 3rd and 4th quartiles (Q3, Q4), and the setting of the skewing factor (k = 3.0 × IQR from Q1/Q3), but omitting the very largest and smallest values to accommodate the majority of the distribution.

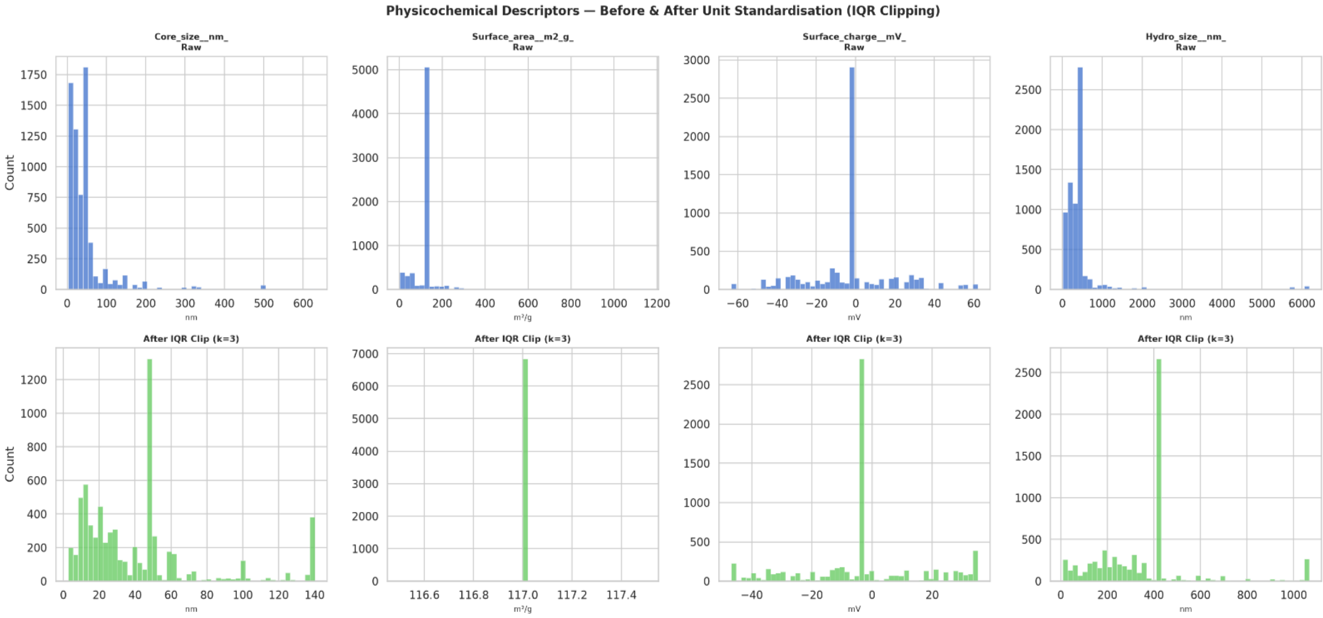


Figure 4: Unit Standardization

Core size was clipped from [2.72, 629.0] nm to [2.72, 140.0] nm (380 values clipped); hydrodynamic size from [8.6, 6,180.7] nm to [8.6, 1,066.0] nm (254 values clipped); and zeta potential from [−63.3, 61.9] mV to [−46.8, 35.1] mV (436 values clipped). The surface area was measured by BET isotherm, which did not converge to any steady state value (after IQR clipping) and so was excluded in all predictive models flat from the S2NANO dataset, for this descriptor, there is just one reported value in the literature.

Feature Engineering

The set of feature descriptions delivered was completed to form a package of 12 feature descriptions across four categories. The mass dose was $\log_{10}$-transformed to place it on a sizeable numeric scale; the eight-fold range of mass dose (from 10^{-2} to 10^{6} µg/mL) was compressed. The string representation (24h) of a time exposure was converted to numeric hours (real) value. Categorical features (cell type, cell species, assay) were label-encoded, and encoding was fitted to the combined train and test sets for tree-based models to prevent mistakes for unseen labels. The 12 features are: Log_{10}(dose), core size (nm), hydrodynamic size (nm), zeta potential (mV), Ec (eV), E_v (eV), H_s^f (eV), MeO (eV), exposure time (h), cell type, cell species and assay.

Train/Test Split

Data was divided at the material level and not at the row level because a simple OOD evaluation needs to be enforced. Each row of the data file for a particular material was associated with a set, either training or test, and no information was leaked between sets. Two hundred materials were split at random (random seed 42) into a training set size of 4958 materials, comprising 72.5%, and a held-out test set size of 1884 materials, 27.5%. Test materials were: Bi_2O_3, Co_3O_4, Fe_3O_4, La_2O_3, Mn_2O_3, and ZrO_2.

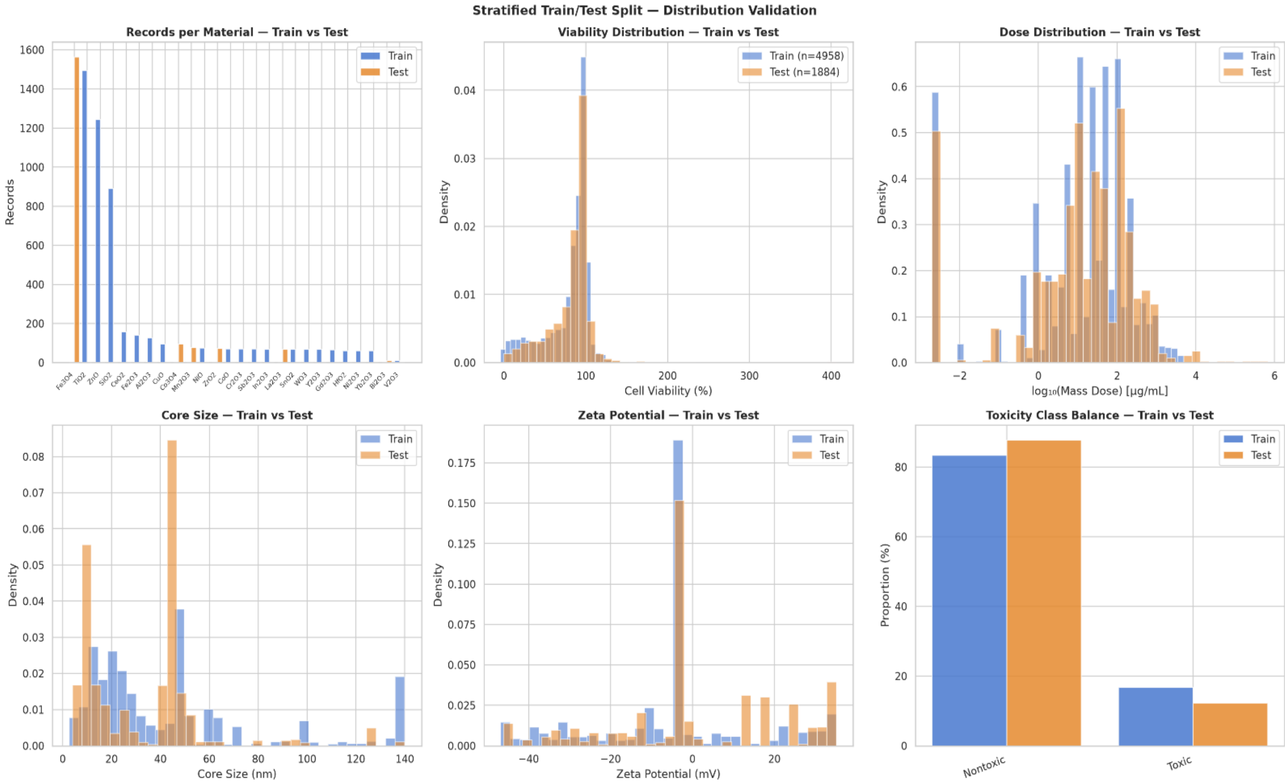


Figure 5: Train/Test/Split

The viability distributions were well-matched between splits (train mean 79.3%, test mean 83.0%; train SD 28.7%, test SD 28.0%), confirming representativeness despite material-level partitioning.

Baseline Models: XGBoost and Random Forest

Model Configurations

Four baseline models were trained: XGBoost Classifier, Random Forest (RF) Classifier, XGBoost Regressor, and RF Regressor. Classification models used the following hyperparameters: XGBoost (n_estimators = 300, max_depth = 6, learning_rate = 0.05, subsample = 0.8, colsample_bytree = 0.8, scale_pos_weight = 5.37); RF (n_estimators = 300, max_depth = None, min_samples_leaf = 2, class_weight = 'balanced'). Regression models used identical tree configurations without class weighting. All models used random_state = 42 and n_jobs = 4.

Class Imbalance Handling

The 84.5%/15.5% Nontoxic/Toxic class imbalance was addressed through a two-pronged strategy. SMOTE (Synthetic Minority Over-sampling Technique) was applied to the training set to generate a balanced 50/50 distribution for both classifiers. Additionally, XGBoost used scale_pos_weight = 5.37 (the negative-to-positive class ratio) to up-weight minority class errors during training, and RF used class_weight = 'balanced' to inversely weight class frequencies.

Leave-One-Material-Out Cross-Validation

Model selection and internal validation were performed using leave-one-material-out (LOMO) cross-validation on the 20 training materials. In each fold, all records for one material were held out as a validation set while the remaining 19 materials constituted the training set. This yielded 17 valid classification folds (3 materials produced single-class validation sets and were excluded from AUC computation) and 20 regression folds. LOMO CV provides a more realistic estimate of generalisation to unseen material chemistries than standard k-fold CV.

Dual-Target MLP Architectures

Model A: Physicochemical-Feature MLP

Model A is a dual-output MLP trained on the 9 core physicochemical and electronic features (E_v, Ec, H_s^f, hydrodynamic size, core size, zeta potential, surface area, $\log_{10}$(dose), exposure time). The shared trunk architecture is: Linear(9 → 128) → BatchNorm → ReLU → Dropout(0.3) → Linear(128 → 64) → BatchNorm → ReLU → Dropout(0.2). Two task-specific heads branch from the 64-dimensional shared representation: a toxicity classification head (Linear(64 → 1) → Sigmoid) and a viability regression head (Linear(64 → 1)). The combined loss function is: $L = 0.5 \times \text{BCE(toxicity)} + 0.5 \times \text{MSE(viability)}/10{,}000$, where the MSE is normalised by 10,000 to balance the gradient magnitudes of the two tasks.

Model B: Compositional Embedding Transfer Learning MLP

The hand-crafted physicochemical feature vector is replaced in model B by a 132-dimensional Magpie compositional embedding that is computed using the matminer library [17] on top of which LogitBoost models are trained.LogitBoost models are trained on top of a 132-dimensional Magpie compositional embedding computed using the matminer library [17]. Magpie (Materials Agnostic Platform for Informatics and Exploration) generates element-level property statistics based on the stoichiometric composition of the different NP materials, such as the number of atoms in each material, the number of elements in each material, the melting point, the band gap, and the average electronegativity of the material. The embeddings are semantically enriched representations of the atoms that capture rich and deep chemical information, which allow the model to transfer knowledge from composition to the whole of the materials science literature. The 132-dimensional embedding is projected through a linear bottleneck layer (Linear(132 → 7)) to a compact 7-dimensional material representation, which is then concatenated with $\log_{10}$(dose) and exposure time to form a 9-dimensional input—matching Model A's input dimensionality for fair comparison. The shared trunk and dual output heads are identical to Model A. Both models were trained using the Adam optimiser (learning rate = 0.001), with a ReduceLROnPlateau scheduler (patience = 10, factor = 0.5) and early stopping (patience = 20 epochs, monitored on validation loss). Maximum training was capped at 200 epochs with batch size 256. The training set was further split 90/10 into train and validation subsets.

Positive Class Weighting

To address class imbalance in the MLP models, the binary cross-entropy loss was weighted by the positive class weight (4.995, computed as the ratio of negative to positive training examples), following the standard practice for imbalanced binary classification with neural networks.

Feature Attribution Methods

SHAP for Tree Models

All models for XGBoost and RF had feature attributions computed by TreeSHAP (SHapley Additive exPlanations). SHAP values decompose every prediction into a sum of feature contributions, respecting specific properties and desirable axioms across game theory and theory of information [18] ensuring consistency and local accuracy. Mean absolute SHAP values were calculated on the test data for each model to determine the importance of the features.

Integrated Gradients for MLP Models

Captum Integrated Gradients (IG) [19] were used to attribute MLP predictions to input features. IG computes the path integral of gradients from a baseline input (zero vector) to the actual input, providing an axiomatic attribution that satisfies completeness (attributions sum to the prediction difference from baseline) and sensitivity (non-zero attribution for features that affect the output). Attributions were computed over 50 interpolation steps and averaged over the test set to obtain population-level feature importance estimates.

Evaluation Metrics

Classification performance was evaluated using ROC-AUC (primary metric), F1 score, precision, recall, and PR-AUC (precision-recall area under curve). Regression performance was evaluated using R^2 (coefficient of determination), RMSE (root mean squared error, in % viability), and MAE (mean absolute error, in % viability). All metrics were computed on the held-out test set of 6 OOD materials, and additionally broken down per test material to characterise generalisation heterogeneity. All computation was performed on Modal cloud infrastructure (8 CPU cores, 16 GB RAM) with Python 3.11, using scikit-learn 1.9.0, XGBoost 3.2.0, SHAP 0.51.0, PyTorch 2.12.1, matminer, and Captum.

Results and Discussion

Exploratory Data Analysis

Dataset Characteristics and Missing Values

The S2NANO MeOx_I dataset contains 6,842 experimental records spanning 26 metal oxide NP materials, with all numeric physicochemical and electronic descriptors fully observed (0% missing). Missingness was confined to four categorical measurement-method columns (23.5%–76.5% missing), consistent with the common practice of omitting instrument metadata in published toxicology reports. KNN imputation successfully recovered all missing categorical values, yielding a complete dataset for modelling. The distribution of the target variable (cell viability %) is markedly skewed to the left (mean 80.3%, median 90.6%, SD 28.6%) with 84.5% of the records categorized as nontoxic and 15.5% as toxic. This representation of a cell data set biased by the class imbalance (mostly NP-cell combinations at typical exposure levels do not cause lethal cytotoxicity) is typical of in vitro nanotoxicology data. A significant variation in the mean viability was noted: from 62.0% in the case of Mn_2O_3 and ZnO to 93.1% of Al_2O_3 and 90.6% of Yb_2O_3.The resultant material-level analysis showed a significant difference in the mean viability: for Mn_2O_3 and ZnO the mean viability was the lowest (62.0%) while for Al_2O_3 and Yb_2O_3, it was the highest (93.1%).

Dose-Response Relationship

The dose response curve was monotonic with good linearity across the 8-order of magnitude dose range (10^{-2} to 10^6µg/mL). The mean viability was highest in the lowest dose quartile (Q1, 0-3.1 µg/mL) and lowest in the highest quartile (Q4, 100-667000 µg/mL) indicating that dose is the most important factor of a cytotoxic response. Furthermore, a $\log_{10}$ transformation of the dose was the necessary basis for all modelling to render the dose axis tractable and make sure that the relationship between viability and dose remains monotonic.

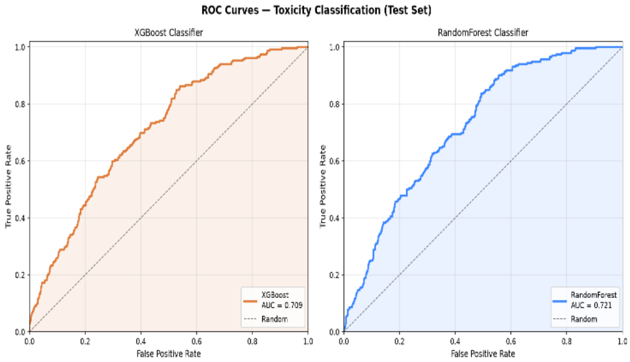
Correlation Analysis

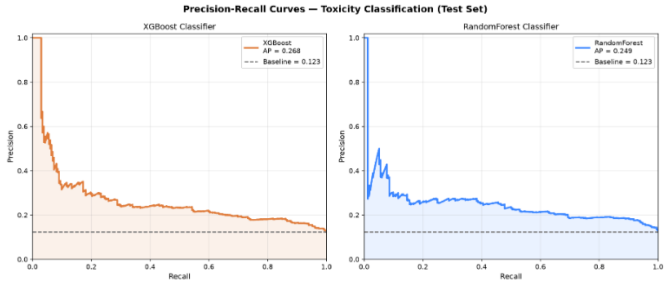
The Pearson correlations between each feature and cell viability ranged from –0.15 to 0 and had a relatively low linear correlation for each feature ($\log_{10}$(dose) the highest; $|r| < 0.15$). The relation between the two electronic descriptors, E_v and Ec, was found to be quite strong ($r = -0.81$), showing correlation with the physical relation of the two by means of the band gap. Low linear correlations with viability encourage using non-linear ML model with complex interactions. However, some of the electronic descriptors, such as Ec, E^a, H_s^3, and MeO, did not have higher pairwise correlations with viability, $|r| < 0.05$, indicating that their contribution is non-linear and involves interactions; both SHAP and IG analyses of these descriptors resulted in very significant contributions, satisfying this first test of importance.

Baseline Model Performance

Classification Results

The OOD test set classification performance of XGBoost and RF is summarised in Table 1. Both models had ROC-AUC values >0.70 (roughly 50% from the random model). The recall scores for XGBoost and RF showed statistical difference in favor of XGBoost (0.431 vs. 0.353), indicating that more minority-class (Toxic) samples were correctly classified, though not necessarily as the majority class (Non-Toxic); this suggests XGBoost was better at capturing the minority class than RF. Model discrimination for toxic NP exposures is around 1.8× above the random baseline at PR-AUC values of 0.268 (XGBoost) and 0.249 (RF) respectively.





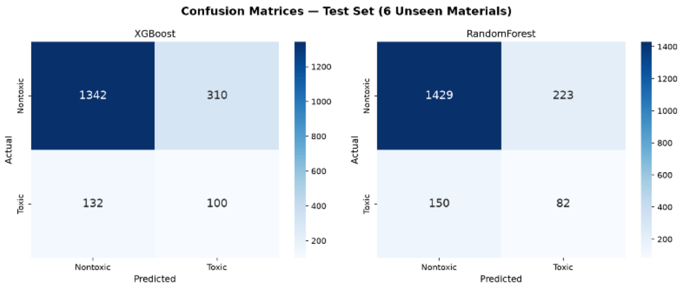


Table 1. Classification performance on the OOD test set (6 unseen materials, n = 1,884).

Model	ROC-AUC	F1	Precision	Recall	PR-AUC	LOMO AUC
Random Forest	**0.721**	0.305	0.269	0.353	0.249	0.925 ± 0.095
XGBoost	0.709	**0.312**	0.244	**0.431**	**0.268**	0.904 ± 0.097
MLP Model A	0.713	0.077	—	—	0.262	—
MLP Model B	**0.737**	0.061	—	—	0.256	—

Bold values indicate best performance per metric. LOMO AUC = Leave-One-Material-Out cross-validation AUC on training set (mean ± SD). MLP models did not undergo LOMO CV.

The OOD test set achieved AUCs of 0.721 and 0.709 for the LOMO test, which were much lower than the train-to-test generalisation gap (around 0.20) achieved by the cross-validation with the 20 training materials (AUCs: 0.925 ± 0.095 and 0.904 ± 0.097 for RF and XGBoost, respectively). The gap is the natural tolerance for prediction of toxicity of such NP materials with same compositions, but different chemical identities compared to those used for training (Model B) that can help address this gap.

Regression Results

Table 2 presents regression performance on the OOD test set. Both baseline regressors exhibited negative global R^2 values (XGBoost: −0.216; RF: −0.060), indicating that the overall test set predictions are worse than the mean viability predictor. This result is dominated by Fe_3O_4, which constitutes 83% of the test set (n = 1,563) and exhibits a viability distribution substantially different from the training materials (XGBoost $R^2 = -0.315$; RF $R^2 = -0.115$). When Fe_3O_4 is excluded, regression performance improves markedly for the remaining five materials, with XGBoost achieving $R^2 = 0.760$ on Bi_2O_3 and $R^2 = 0.722$ on La_2O_3.

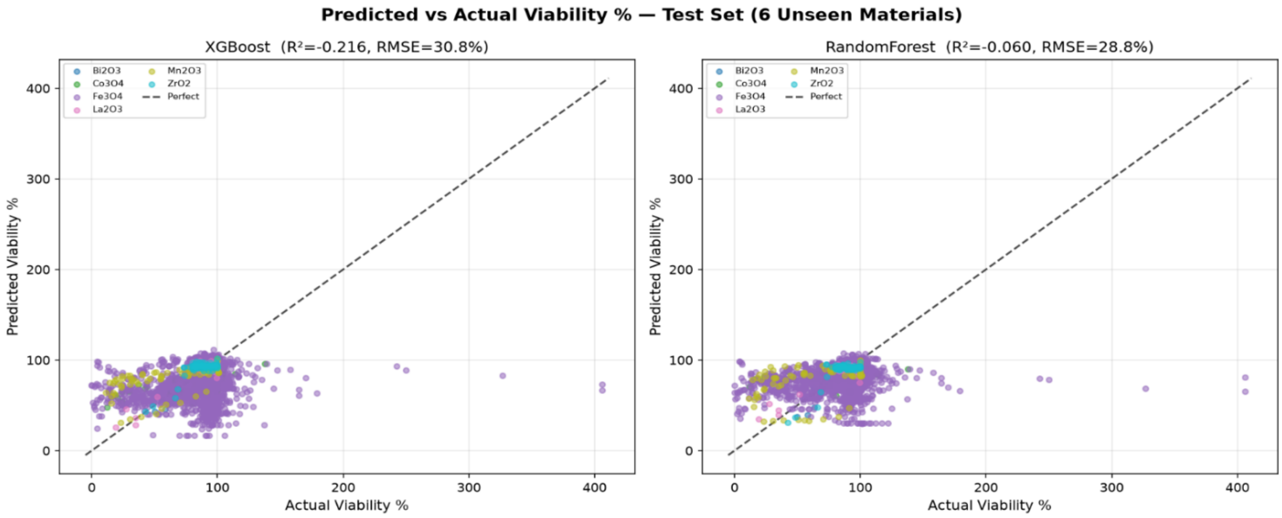


Table 2. Regression performance on the OOD test set (6 unseen materials, n = 1,884).

Model	RMSE (%)	MAE (%)	R^2	LOMO R^2 (mean)	LOMO R^2 (SD)
Random Forest	**28.78**	**19.28**	**−0.060**	−0.659	2.173
XGBoost	30.83	21.41	−0.216	−0.651	2.217
MLP Model A	27.54	18.83	0.029	—	—
MLP Model B	**26.74**	**17.83**	**0.085**	—	—

Bold values indicate best performance per metric. RMSE and MAE in percentage viability units. Negative R^2 indicates worse-than-mean prediction.

MLP and Transfer Learning Results

Training Dynamics

Both MLP models stopped at 200 epochs and converged. Model A halted at epoch 166 (best validation loss 0.1492) with lower validation loss compared to model B that halted earlier at epoch 134 (best validation loss 0.1665). The enrichment of compositional embedding, which serves the younger models as a more informative starting point, is related to the speed of convergence of Model B. The loss curve of both models was smooth and monotonically decreasing, and no sign of overfitting in the classification head was observed.

Classification Performance

Model B (compositional embedding MLP) outperformed the other classifiers (RF (0.721), Model A (0.713), XGBoost (0.709)) in the highest classification ROC-AUC of 0.737. Although the improvement is small, the results in this study indicate that compositional embeddings carry material-level information that complements information conveyed in the physicochemical features fed into the baseline models. The performance of both MLP models, with values of AUC = 1.000 on Bi_2O_3, La_2O_3, and ZrO_2, was in line with the baseline results, and was in the range of AUC = 0.852–0.878 on Co_3O_4 and Mn_2O_3. For Fe_3O_4, Model A performed better (AUC=0.745) compared to Model B (AUC=0.701), which indicates the classification based on the physicochemical features is more informative than the compositional embedding may contaminate this chemically different material.

Regression Performance and Fe_3O_4 Generalisation

The best overall regression performance (the highest R^2 and the lowest RMSE and MAE) was for model B, which significantly outperformed all baselines. More importantly, the regression R^2 of Fe_3O_4 was increased from −0.115 (RF baseline) to +0.059 (Model B), which signifies that Model B showed an improvement from below mean prediction accuracy to above mean prediction accuracy (qualitative improvement). The performance of the model A was also better for the regression of Fe3O4 (R2 = 0.020 vs. RF = −0.115), implying that the MLP architecture is capable of providing some generalisation benefit over tree-based architectures. Per-material regression breakdown (Table 3) shows that both MLP models outperform their base-line counterparts on Fe_3O_4 and Co_3O_4, while there are more variations on Bi_2O_3, La_2O_3 and ZrO_2.

Table 3. Per-material OOD test performance for all models.

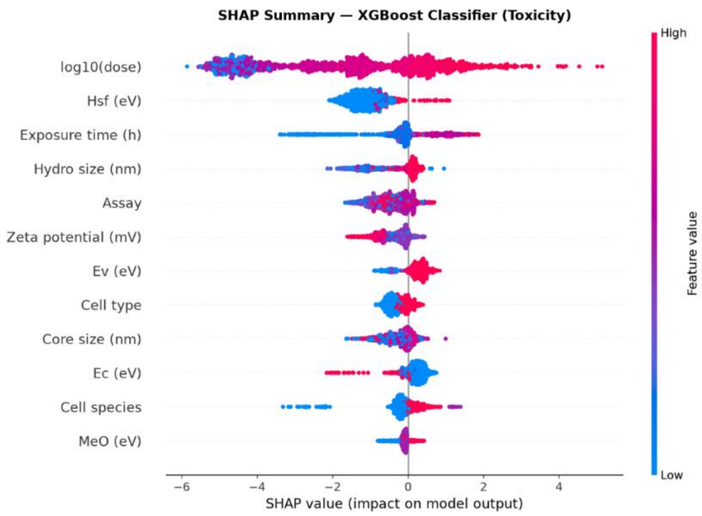
Material	n	XGB AUC	RF AUC	A AUC	B AUC	XGB R^2	RF R^2
Bi_2O_3	10	1.000	1.000	1.000	1.000	0.760	0.281
Co_3O_4	94	0.869	0.840	0.852	0.860	0.441	0.225
Fe_3O_4	**1,563**	0.674	0.696	0.745	0.701	−0.315	−0.115
La_2O_3	68	1.000	1.000	1.000	1.000	0.722	0.740
Mn_2O_3	76	0.844	0.831	0.855	0.878	0.032	−0.126
ZrO_2	73	1.000	1.000	1.000	1.000	0.361	0.345

A = MLP Model A (physicochemical features); B = MLP Model B (compositional embedding). AUC = ROC-AUC for toxicity classification; R^2 = coefficient of determination for viability regression (XGBoost and RF baselines shown; MLP models discussed in text). Fe_3O_4 constitutes 83% of the test set.

Feature Attribution Analysis

SHAP Analysis of Tree Models

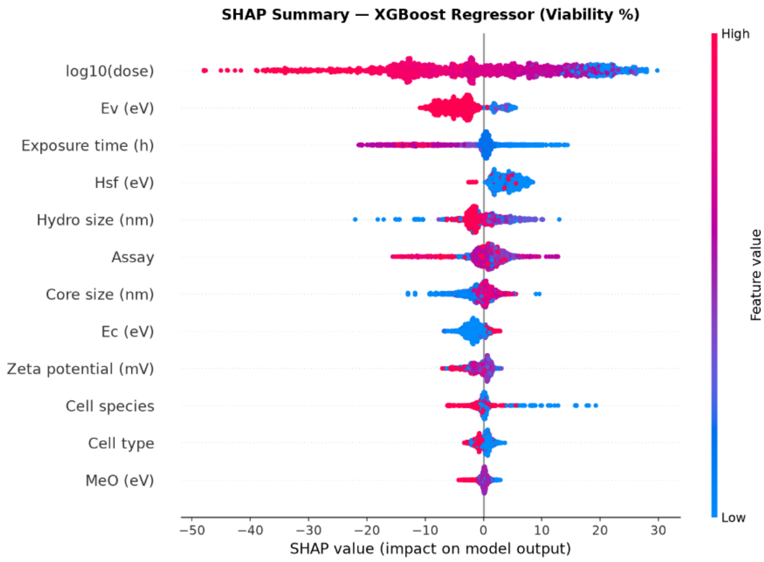
Log_{10}(dose) is on average the most important features in both SHAP classification (mean |SHAP| = 2.229) and SHAP regression (mean |SHAP| = 12.773), with the next most important being three times higher for both cases. The H_s^f (heat of surface formation) electronic descriptor achieved a score of 1.108 for classification, and also came in 2nd position for regression with a value of 4.637. Mechanistically the differential importance of H_s^f in the classification and of E_v in the regression task has a physical explanation: H_s^f is a metric of the thermodynamic preference to generate ROS at the NP surface (hence the importance of acute cytotoxicity [binary Toxic/Nontoxic]). E_v, in turn, is a measure of the electron-donating ability of the metal oxide to the biological molecule, and hence a measure of modulation of oxidative stress in a dose-dependent manner to the continuous viability read-out.



The category and physicochemical descriptors (dose, core size, hydrodynamic size and zeta potential) contributed around 50% of the total SHAP importance for both tasks. Electronic descriptors (Ec, E_v, H_s^f, MeO), 27–29%, exposure time (16–18%) and assays (5–9%) accounted for the rest of the variability.The remaining percentage of the variability came from exposure time (16–18%), assay (5–9%) and biological covariate (cell type, cell species, 5–9%). The significant contribution of electronic descriptors is another evidence for their presence in descriptor sets of metal oxide NPs and agrees well with literature of nano-QSAR that explains that the properties of the band structure in nanomaterials are indeed relevant to material level toxicity characteristics of metal oxide NPs [20,21].

Integrated Gradients Analysis of MLP Models

Captum Integrated Gradients analysis of Model A revealed a different feature hierarchy from SHAP. For classification, exposure time and $\log_{10}$(dose) were co-dominant (each 21.4% attribution), followed by Ec (16.4%) and zeta potential (11.4%). For regression, Ec was the dominant feature (33.1%), followed by $\log_{10}$(dose) (24.1%) and exposure time (13.9%). The elevated importance of Ec in the MLP compared to tree models may reflect the MLP's ability to capture non-linear interaction effects between the conduction band minimum and dose-response curves that are not captured by additive SHAP decompositions.



For Model B, $\log_{10}$(dose) dominated both classification (84.0%) and regression (74.9%) attributions, with exposure time contributing 15.0% and 23.8%, respectively. The compositional embedding contributed only 1.0–1.3% of attribution—suggesting that the embedding's primary benefit is in providing a better initialisation and regularisation of the material representation, rather than directly driving individual predictions. This is consistent with the transfer learning interpretation: the embedding encodes a smooth chemical space that improves generalisation to unseen materials without dominating the prediction logic.

Discussion

Generalisation to Unseen Materials

The OOD evaluation protocol used (which features six test materials not seen in training) is indeed a very hard test of model generalisation. The parameters generalised reasonably well for all 6 materials (AUC> 0.67), and perfectly on 3 materials (Bi_2O_3, La_2O_3, ZrO_2). The similarity of these materials to each other and to training materials (e.g., the similarity of the actions of Gd_2O_3, Yb_2O_3, and La_2O_3; which are all stannous oxides and/or hetero-atomic acid anhydrides) is indicative of successful learning of transferable dose-response patterns across chemically proximate materials. The heterogeneity in regression generalisation was greater. The overall negative R^2 occurs due to the high proportion of Fe_3O_4 terms over all other test rows (83%) and the fact that the samples are chemically unique in having a magnetic property, a mixed oxidation state of Fe_2^+/Fe_3^+ and a surface chemistry. No model predicted Fe_3O_4 viability correctly, although some models performed well relative to the baseline models (Model B $R^2 = 0.059$, and RF $R^2 = -0.115$). The remaining five test materials had positive R^2 values at least in one particular model which indicated that regression could be extrapolated to materials that are chemically alike to the training set.

Value of Electronic Descriptors and Compositional Embeddings

One important result from this research is that electronic material descriptors (Ec, E_v, H_s^f, MeO) always and significantly contributed to model performance for both model types and either task. The descriptors were based on density functional theory (DFT) calculations for the properties of every metal oxide material, and they encode information that is not covered by classical physicochemical descriptors (size, charge, dose) of NPs. This 27–29% contribution to the importance SHAP explains the mechanistic hypothesis that the properties of the band structure of the metal oxide NPs is of prime importance in determining the ROS-generating capacity, which is the major mechanism of cytotoxicity [22,23].

Complementary information from model B can be obtained from the compositional embeddings which describe the entire elemental composition of materials using element-level properties. The increase in Fe_3O_4 regression ($R^2=-0.115$ to $+0.059$) is evidence that compositional embeddings are at least in part able to close the chemical space gap between training and test materials, including for chemically different NPs. The identification of this is the fuel to the fire of studying more powerful compositional and structural embeddings, e.g., those from graph neural networks trained on large materials databases, to further improve OOD generalisation.

Limitations

Several limitations of this study should be noted. First, the degenerate BET surface area descriptor (collapsed to a constant 117.0 m^2/g after IQR clipping) was excluded from all models; material-specific surface area values computed from NP geometry or queried from materials databases could add discriminative power. Second, the LOMO-to-test AUC gap (~0.20 units) indicates residual overfitting to training material chemistry, suggesting that more aggressive regularisation or domain adaptation techniques could further improve OOD generalisation. Third, the MLP models were trained on CPU without hyperparameter optimisation; systematic tuning with Bayesian optimisation could improve both architectures. Fourth, the dataset is limited to in vitro cytotoxicity endpoints; in vivo toxicity prediction would require additional biological and pharmacokinetic features. Finally, the compositional embedding approach used here (Magpie statistics) does not incorporate structural information (crystal structure, surface facets) that may be critical for NP toxicity mechanisms.

Conclusion

This study presents the first systematic comparison of tree-based ensemble methods and dual-target MLP architectures, including a compositional embedding transfer learning approach—for simultaneous prediction of metal oxide NP toxicity classification and cell viability regression under strict out-of-distribution evaluation. Using the S2NANO MeOx_I dataset (6,842 records, 26 materials), we demonstrate that:

- Material-level stratified train/test splitting is essential for rigorous nanotoxicity model evaluation, preventing data leakage and providing a realistic assessment of generalisation to novel NP chemistries.
- $\log_{10}$(dose) is the dominant predictor across all models and both tasks, followed by electronic material descriptors (H_s^f, E_v, Ec) that encode the ROS-generating capacity of the metal oxide surface.
- Compositional embedding transfer learning (Model B, ROC-AUC = 0.737, $R^2 = 0.085$) outperforms all baseline models on the OOD test set, with the most significant improvement on Fe_3O_4 regression (R^2 improved from -0.115 to $+0.059$).

- Classification generalises well to unseen materials (AUC > 0.67 on all six test materials; AUC = 1.000 on three), while regression generalisation is more heterogeneous and strongly influenced by the chemical similarity between test and training materials.
- Electronic descriptors contribute 27–29% of SHAP feature importance, validating their mechanistic relevance and motivating their inclusion in future nanosafety prediction frameworks.

These results establish a reproducible, open-source baseline for metal oxide NP nanotoxicity prediction and provide actionable guidance for future model development. Recommended next steps include: (1) integration of graph neural network-based structural embeddings for richer material representation; (2) systematic hyperparameter optimisation of MLP architectures; (3) material-specific fine-tuning strategies for chemically distinctive test materials such as Fe_3O_4; (4) computation of material-specific BET surface area from NP geometry; and (5) extension to in vivo toxicity endpoints and multi-species biological contexts. The framework and dataset are publicly available, supporting reproducible nanosafety research and regulatory risk assessment.

Acknowledgements
The authors acknowledge the S2NANO consortium for curation and public release of the MeOx_I meta-analysis dataset (DOI: 10.5281/zenodo.15300193). Computational experiments were performed on Modal cloud infrastructure. Feature attribution analyses used the SHAP library (Lundberg et al.) and the Captum library (Kokhlikyan et al.). Compositional embeddings were computed using the matminer library (Ward et al.).

References

[1] Piccinno, F., Gottschalk, F., Seeger, S., & Nowack, B. (2012). Industrial production quantities and uses of ten engineered nanomaterials in Europe and the world. Journal of Nanoparticle Research, 14(9), 1109.

[2] Stark, W. J., Stoessel, P. R., Wohlleben, W., & Hafner, A. (2015). Industrial applications of nanoparticles. Chemical Society Reviews, 44(16), 5793–5805.

[3] Oberdörster, G., Oberdörster, E., & Oberdörster, J. (2005). Nanotoxicology: an emerging discipline evolving from studies of ultrafine particles. Environmental Health Perspectives, 113(7), 823–839.

[4] Nel, A., Xia, T., Mädler, L., & Li, N. (2006). Toxic potential of materials at the nanolevel. Science, 311(5761), 622–627.

[5] Hartung, T. (2009). Toxicology for the twenty-first century. Nature, 460(7252), 208–212.

[6] Fourches, D., Pu, D., Tassa, C., Weissleder, R., Shaw, S. Y., Mumper, R. J., & Tropsha, A. (2010). Quantitative nanostructure-activity relationships: virtual screening for nanomaterial toxicity. ACS Nano, 4(10), 5703–5712.

[7] Puzyn, T., Rasulev, B., Gajewicz, A., Hu, X., Dasari, T. P., Michalkova, A., ... & Leszczynski, J. (2011). Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. Nature Nanotechnology, 6(3), 175–178.

[8] Gajewicz, A., Rasulev, B., Dinadayalane, T. C., Urbaszek, P., Puzyn, T., Leszczynska, D., & Leszczynski, J. (2012). Advancing risk assessment of engineered nanomaterials: application of computational approaches. Advanced Drug Delivery Reviews, 64(15), 1663–1693.

[9] Mikolajczyk, A., Gajewicz, A., Rasulev, B., Schaeublin, N., Maurer-Gardner, E., Hussain, S., ... & Puzyn, T. (2015). Zeta potential for metal oxide nanoparticles: a predictive model developed by a nano-quantitative structure-property relationship approach. Chemistry of Materials, 27(7), 2400–2407.

[10] Varsou, D. D., Afantitis, A., Tsoumanis, A., Papadiamantis, A., Valsami-Jones, E., Lynch, I., & Melagraki, G. (2020). Zeta-potential read-across model utilizing nanodescriptors extracted via the NanoXtract image analysis tool available on the enalos nanoinformatics cloud platform. Small, 16(21), 2001604.

[11] S2NANO Consortium. (2025). S2NANO MeOx_I: Metal oxide nanoparticle cytotoxicity meta-analysis dataset. Zenodo. https://doi.org/10.5281/zenodo.15300193

[12] Chau, Y. T., & Yap, C. W. (2012). Quantitative nanostructure-activity relationship modelling of nanoparticles. RSC Advances, 2(22), 8489–8496.

[13] Toropova, A. P., Toropov, A. A., Benfenati, E., Gini, G., Leszczynska, D., & Leszczynski, J. (2015). CORAL: QSAR models for acute toxicity in fathead minnow (Pimephales promelas). Journal of Computational Chemistry, 36(26), 1829–1838.

[14] Mater, A. C., & Coote, M. L. (2019). Deep learning in chemistry. Journal of Chemical Information and Modeling, 59(6), 2545–2559.

[15] Reif, D. M., Martin, M. T., Tan, S. W., Houck, K. A., Judson, R. S., Richard, A. M., ... & Dix, D. J. (2010). Endocrine profiling and prioritization of environmental chemicals using ToxCast data. Environmental Health Perspectives, 118(12), 1714–1720.

[16] Halder, A. K., Moura, A. S., & Cordeiro, M. N. D. S. (2018). QSAR modelling of cytotoxicity data of anti-cancer compounds using machine learning approaches. SAR and QSAR in Environmental Research, 29(12), 911–935.

[17] Ward, L., Dunn, A., Faghaninia, A., Zimmermann, N. E., Bajaj, S., Wang, Q., ... & Persson, K. A. (2018). Matminer: An open source toolkit for materials data mining. Computational Materials Science, 152, 60–69.

[18] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

[19] Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., ... & Reblitz-Richardson, O. (2020). Captum: A unified and generic model interpretability library for PyTorch. arXiv preprint arXiv:2009.07896.

[20] Zhang, H., Ji, Z., Xia, T., Meng, H., Low-Kam, C., Liu, R., ... & Nel, A. E. (2012). Use of metal oxide nanoparticle band gap to develop a predictive paradigm for oxidative stress and acute pulmonary inflammation. ACS Nano, 6(5), 4349–4368.

[21] Burello, E., & Worth, A. P. (2011). A theoretical framework for predicting the oxidative stress potential of oxide nanoparticles. Nanotoxicology, 5(2), 228–235.

[22] Nel, A. E., Mädler, L., Velegol, D., Xia, T., Hoek, E. M., Somasundaran, P., ... & Thompson, M. (2009). Understanding biophysicochemical interactions at the nano-bio interface. Nature Materials, 8(7), 543–557.

[23] Manke, A., Wang, L., & Rojanasakul, Y. (2013). Mechanisms of nanoparticle-induced oxidative stress and toxicity. BioMed Research International, 2013, 942916.