KONA Powder and Particle Journal

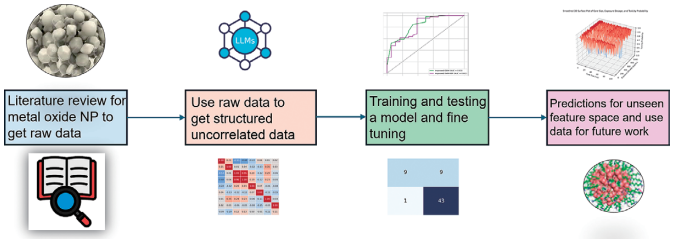
KONA Powder and Particle Journal No. 43 (2026) 238–248/https://doi.org/10.14356/kona.2026019 Original Research Paper

Data Curation to Develop Machine Learning Models for Assessing the Toxicity of Nanoparticles†

Soham Savarkar, Jason Gibson, Vasanthakumar Balasubramanian, Brij Moudgil* and Richard Hennig

Department of Materials Science and Engineering, University of Florida, USA

Metal oxide nanoparticles (NPs) are extensively employed in the biomedical, environmental, and industrial domains due to their unique physicochemical properties. However, concerns regarding their potential cytotoxicity require the development of accurate predictive models to assess nanoparticle safety. In this study, we present a machine learning-based framework for predicting the toxicity of metal oxide NPs using curated physicochemical descriptors. Data were systematically extracted and structured from 140 peer-reviewed publications, focusing on four representative metal oxide nanoparticles (ZnO, AgO, CuO, SiO_2). To ensure accessibility and consistency, the dataset was structured using a Large Language Model (LLM) API and designed to be well-balanced and minimally correlated. The maintenance of a low correlation between features (average Pearson correlation = 0.19) was prioritized to reduce redundancy and improve the interpretability of the model results. Feature selection and Principal Component Analysis (PCA) confirmed that a subset of physical descriptors effectively captured toxicity-related trends. The optimized Gradient Boosting Machine (GBM) and Support Vector Machine (SVM) models achieved predictive accuracies of 77 % and 81 %, respectively, without overfitting. In addition, a synthetic dataset was generated to investigate the joint effects of core size and exposure dosage on toxicity probability. Overall, this study aims to provide a predictive approach framework for nanotoxicity assessment that might offer guidance for the rational design of safer nanoparticles.



Keywords: nanoparticle, cytotoxicity, correlation, machine learning, gradient boosting machine, support vector machine

1. Introduction

Metal oxide nanoparticles, including zinc oxide (ZnO), silver oxide (AgO), copper oxide (CuO), and silicon dioxide (SiO_2), have gained considerable interest in biomedical applications due to their distinct physicochemical properties. These nanoparticles demonstrate potent antibacterial, anticancer, and drug-delivery capabilities, making them promising candidates for biosensing, imaging, and therapeutic advancements (Nikolova and Chavali, 2020). Zinc oxide nanoparticles possess intrinsic antimicrobial properties and are widely used in wound healing, bioimaging, and targeted drug delivery (Mishra et al., 2017). However, concerns regarding their toxicity under various biological conditions have hindered achieving the full potential of their biomedical application.

The toxicological behavior of metal nanoparticles is highly dependent on factors such as size, surface chemistry, dissolution rate, and exposure conditions (Vanivska et al., 2024). Nanoparticles can induce oxidative stress, DNA damage, and inflammatory responses, raising safety concerns for clinical applications (Manke et al., 2013). Traditional toxicity assessments rely on in vitro and in vivo models, which provide valuable insights but face challenges such as labor-intensive procedures, prolonged experimental periods, and high costs (Madorran et al., 2019). Reproducibility issues (Savage et al., 2019) and ethical concerns (Krewski et al., 2010) further complicate toxicity assessment.

Moreover, protein corona formation, aggregation, and environmental conditions influence nanoparticle interactions with biological systems, making empirical predictions difficult (Saptarshi et al., 2013). Since the early 2000s, Machine Learning (ML) has emerged as a powerful tool for toxicity prediction, enabling rapid data processing and pattern identification that traditional statistical methods may overlook (Alowais et al., 2023). Researchers continue investigating various ML models to determine the most influential physicochemical properties that contribute to toxicity, helping to develop safer nanoparticle designs (Zhang et al., 2022). ML-based classification models have achieved high accuracy in predicting metal oxide nanoparticles toxicity using physicochemical descriptors (Xiao et al., 2024). Despite the advancements in ML-based toxicity modeling, several research gaps remain. Existing datasets

† Received 11 April 2025; Accepted 12 May 2025
J-STAGE Advance published online 23 August 2025
* Corresponding author: Brij M. Moudgil;
Add: Gainesville, FL 32611, USA
E-mail: moudgil@ufl.edu
TEL: +1-352-328-7292 (M) FAX: +1-352-392-7219

are often small, incomplete, or derived from heterogeneous sources, limiting model generalizability (von Borries et al., 2023). Additionally, a lack of parameter standardization leads to variations in reported features (Muralidharan et al., 2024). In the field of machine learning-based modeling research, the reliability and accessibility of datasets play pivotal roles in the development and validation of predictive models. However, it is noted that several research groups either do not declare their datasets or rely heavily on restricted, non-public datasets, thereby limiting the reproducibility and validation of their findings by the broader scientific community. Furthermore, a standard limitation observed in these studies is the absence of diversity in the dataset features, which may lead to overly optimistic and not generalizable models across different experimental conditions.

To address these shortcomings, our study introduces a comprehensive dataset compiled de novo, featuring a wide array of the physical and chemical properties of the metal oxide nanoparticles. By employing advanced ML techniques such as Gradient Boosting Machines (GBM) and Support Vector Machines (SVM), we aim to identify features that would lead to guidelines for designing safer nanoparticles. An objective of this paper is to inform other groups with a structured prompt engineered for making similar datasets. We have included the complete structure of the prompt that can be given for the smooth data extraction structuring and querying in the J-STAGE Data (https://doi.org/10.50931/data.kona.29672717).

2. Materials and methods

2.1 Data extraction and structuring for the dataset

Based on our literature survey and preliminary tests on existing open-source datasets, we created our methodology to curate a dataset. On examining the NanoPharos dataset and its use in reported papers by Nikolova and Chavali (2020) and Zouraris et al. (2025), we conducted some preliminary tests on the NanoPharos dataset. We found that after pre-processing the metal oxide cytotoxicity data (as discussed in detail in **Section 2.3**), the average Pearson correlation for the features of interest was 0.39. A Pearson correlation of 0.39 indicates only a weak-to-moderate linear relationship between the selected features and cytotoxicity. This relatively low correlation indicates that the features lack sufficient predictive power on their own, highlighting the limitations of the existing dataset for building robust machine learning models. While individual features demonstrated low pairwise correlation with cytotoxicity, we also observed redundancy among the features themselves, with several exhibiting high inter-feature correlation. This multicollinearity can obscure the model interpretability and degrade the predictive robustness. To address the challenge of high correlation between features existing in open-source nanotoxicity datasets, such as NanoPharos, we curated a dataset with improved feature distribution and lower inter-feature correlation. For open-source datasets such as NanoPharos, we observed that the variance among features for the key physical parameters of interest was lower than 1. The overall variance increased due to a few features, such as hydrodynamic size, which contains several outlier values differing by orders of magnitude. While these values are not necessarily impractical, such variability poses challenges when training a generalizable machine learning model. Such extreme variability, especially when concentrated in a few features, can distort learning algorithms by introducing scale imbalances and reducing the overall effectiveness of feature generalization. Therefore, a curated dataset with controlled variance across features is essential for training models that can generalize well across diverse nanoparticle types. To overcome these challenges, we developed a structured AI-assisted methodology that automates data extraction, standardization, and querying. This methodology ensures improved dataset reliability, reduces manual errors, and provides a reproducible framework for curating high-quality cytotoxicity datasets for metal oxide nanoparticles. We developed a workflow to extract data from various resources to curate our dataset with the necessary features. A relatively simple yet effective approach was designed using a combination of the OpenAI API and OpenAI's ChatGPT-4o model.

The first step of this process involved identifying relevant target searches on Google Scholar and the Web of Science. Papers explicitly focused on cytotoxicity studies for the metal oxide nanoparticles of interest were chosen. We selected a small subset of metal oxide nanoparticles based on their suitability for building a dataset. The choice of metal oxide nanoparticles was based on (i) abundance of available literature, (ii) wide industrial and medical applications, (iii) variability in properties and (iv) complexity in feature relationships (Palanivinayagam and Damasevicius, 2023), which makes them an ideal starting point. We worked with a total of 140 articles covering four metal oxide nanoparticles in our study: ZnO (NP), AgO (NP), CuO (NP), and SiO_2 (NP). After selecting the target nanoparticles, we proceeded with data extraction. To save time instead of manually retrieving data from the papers, we extracted text-queried data from the PDFs and then analyzed it using the API key. We used the PyMuPDF and the pdfplumber libraries to extract and structure the text and images. The extracted data, along with the corresponding PDF files, were then queried using ChatGPT-4o to identify specific feature values. This combined approach was necessary for two reasons: (1) capturing data presented in graphs and images and (2) using extracted text along with manually passed context prompts to avoid incorrect or false values. After extracting the dataset, random query prompts were used to test the reliability of the dataset on 10 % of

randomly selected entries, which resulted in an accuracy of 96 %. This validation step was essential because large language models can sometimes produce hallucinated or non-existent data. We checked the responses against the reference files and ensured that the LLM-generated values were consistent with the original sources. We employed OpenAI's ChatGPT-4o model to automate feature extraction and ensure consistency in data structuring, thereby reducing manual errors in parsing the nanoparticle properties. One of the significant challenges in curating such a dataset is the absence of standardized data reporting practices and consistent precision across research publications. Therefore, several assumptions and generalizations were embedded into the prompts to ensure that the numerical features were reported as integers and the categorical features were as specific as possible. For example, if the article reported a core size for a nanoparticle as a range (e.g., 20–40 nm), the large language model used the mean value as the dataset entry. Similarly, missing numerical features were estimated using the median value of the respective column as reported previously (Papadiamantis et al., 2021).

2.2 Feature selection

Pertinent feature selection is pivotal in constructing a dataset that not only captures the essence of the underlying phenomena but also enhances the predictive power of the model. This section delineates our criteria and methodology for selecting the most informative features, emphasizing the importance of precision in feature choice to ensure our dataset's relevance and robustness. First, we reviewed the existing literature to identify key physical features such as the core size of the nanoparticle (in nm), the shape of the nanoparticle, the surface area (m^2/g), aggregation state, dissolution rate (mg/L), metal ion release (mg/L), and surface chemistry that characterize a nanoparticle system, focusing on those with established or potential roles in inducing toxicity (Nel et al., 2009). The core size determines the cellular uptake efficiency and intracellular distribution, with smaller particles often showing greater reactivity and deeper tissue penetration. The instances of a core size value less than 100 nm were chosen for our study. Morphology affects how nanoparticles interact with cellular membranes and organelles; for instance, rod-shaped particles may induce higher stress responses than spherical ones. The surface area is directly related to the extent of contact with biological systems and thus governs the rate of surface-driven reactions. The aggregation state influences the effective surface area and bioavailability, thereby modulating the toxicity profiles in vitro and in vivo. The dissolution rate is particularly crucial for metal oxide nanoparticles, as the release of ions contributes to oxidative stress. Metal ion release is a key mechanism of toxicity, especially for ZnO, AgO, and CuO nanoparticles, where ions disrupt cellular homeostasis. Finally, surface chemistry determines protein corona formation, cellular recognition, and immunogenicity, making it a critical modulator of the nanoparticle's biological identity (Sun et al., 2024). Additionally, we have included features that are commonly measured as indicators of nanoparticle toxicity, distinguishing between properties that contribute to toxicity and those that are used to assess its effects, and included them in our dataset. These features are Reactive Oxygen Species (ROS) production, zeta potential, membrane damage, apoptosis, necrosis, IC50, and cell viability, all of which are commonly used indicators of nanoparticle toxicity. ROS production reflects oxidative stress, a primary mechanism by which many nanoparticles induce cellular damage. The zeta potential is a proxy for the surface charge, influencing the nanoparticle–cell interactions, the stability of the nanoparticles in biological media, and the potential for their membrane disruption. Membrane damage, along with apoptosis and necrosis, directly measures cell death pathways triggered by toxic insults. IC50 (half-maximal inhibitory concentration) quantifies the effective dose at which 50 % of the cell population is inhibited, serving as a standardized toxicity threshold. Cell viability captures the overall survival rate of cells and is one of the most widely used endpoints in nanotoxicology (Liu et al., 2015).

We included three additional features in our dataset: the exposure dose of the nanoparticle, the exposure time, and the cell type, which are necessary for the predictions. Finally, we added four fundamental features of the nanoparticle core to our dataset. We chose the group number, period number, atomic weight of the metal in the metal oxide, and the electronegativity difference of the elements in the nanoparticle core to explore how fundamental periodic trends and electronic properties influence the nanoparticles' behavior in biological systems. These properties can affect key interactions such as ion release, oxidative stress potential, and binding affinity to biomolecules, all of which play a vital role in determining nanoparticle toxicity. In total, our curated dataset comprises 20 features, encompassing physical characteristics, toxicity indicators, exposure conditions, and core material properties, each selected to provide a comprehensive and mechanistically relevant basis for the predictive modeling of nanoparticle cytotoxicity. We selected a binary classification system for our toxicity label, where 0 is non-toxic and 1 is toxic. The binary classification of toxicity simplifies the output, enhancing practical applications in safety assessments. To develop a dataset suitable for predictive modeling, it was necessary to assign toxicity labels to each nanoparticle instance based on well-defined criteria. We labeled the data by either setting thresholds for toxicity based on established scientific literature or directly using experimentally reported toxicity classifications when available. For instance, IC50 values below a certain threshold indicate high toxicity. Similarly, cell viability percentages and ROS production levels were

assessed against pre-defined toxicity cutoffs to ensure consistency in classification. To classify an experimental instance as toxic or non-toxic, we decided to determine our label using established thresholds and expert guidance. The individual thresholds were as follows: IC50 value ≤ 100 μg/mL, cell viability ≤ 70 % (as per ISO 10993-5: 2009 guidelines, https://www.iso.org/standard/36406.html), ROS production greater than twice the control value (Nel et al., 2009), or zeta potential in the range between –30 mV and 30 mV (Fröhlich, 2012). These thresholds were considered highly probable indicators of cytotoxicity and were labeled as toxic. For this study, if the cells undergoing apoptosis and necrosis increased by 20 %, we assumed that there was sufficient oxidative stress damage that the system could not recover, and we labeled the nanoparticle system as cytotoxic (Akter et al., 2018). These thresholds were introduced to accommodate the differences in the chemical composition of the metal oxide nanoparticles being used.

2.3 Dataset preparation and correlation analysis

An important step in machine learning is ensuring that our dataset is suitable for training a model. The motivation for developing a methodology to prepare a dataset stems from the scarcity of uncorrelated features reported in the literature (Furxhi et al., 2020). We use a correlation matrix to verify whether a dataset is correlated. A correlation matrix quantifies the linear relationships between the numerical features in a dataset. It is an $n \times n$ symmetric matrix R, where each element R_{ij} represents the Pearson correlation coefficient between features X_i and X_j. The Pearson correlation coefficient was calculated using **Eqn. (1)**.

$$R_{ij} = \frac{\mathrm{cov}(X_i, X_j)}{\sigma_{X_i} \cdot \sigma_{X_j}} \quad (1)$$

where, $\mathrm{cov}(X_i, X_j)$ is the covariance between features X_i and X_j, and σ_{xi} and σ_{xj} are their standard deviations. The Pearson correlation coefficient values range from –1 (indicating an inversely proportional relationship) to +1 (indicating a directly proportional relationship), with 0 indicating no linear relationship between any two features. This matrix provides insights into feature dependencies, aiding in relevant feature selection and multicollinearity analysis.

Before computing the covariance matrix, we preprocessed the data by removing the rows with missing values and encoding the categorical features. Categorical data were label-encoded, except for the toxicity label, which was assigned a binary encoding. The feature for cell type, which consists of over 25 different cell types, was categorized into five major groups:

- 0: Human Cancer Cells
- 1: Normal Human Cells
- 2: Mouse/Rat Cells
- 3: Non-Mammalian Cells
- 4: 3D Stem Cells

Similarly, features such as surface chemistry and ROS production score have been categorized based on their scores and have been provided in the J-STAGE Data (https://doi.org/10.50931/data.kona.29672717). We have classified the input and output features as listed in **Table 1** for our analysis. In addition, we plotted the physical features, toxicity indicators, exposure features, and fundamental features in our dataset for the correlation matrix of the nanoparticles (see **Fig. 1**). The correlation matrix, shown in **Fig. 1**, illustrates the relationships among all features. The color intensity indicates the strength and direction of the correlation, where blue represents negative correlations, red represents positive correlations, and white represents

Table 1 Classification of the input and output features for nanoparticle toxicity modeling.

Input variables (nanoparticle characteristics and experimental conditions)	Output variables (performance or biological responses)
core size (nm)	ROS production
zeta potential (mV)	membrane damage (%)
surface area (m^2/g)	apoptosis (%)
aggregation state	necrosis (%)
dissolution rate (mg/L)	IC50 value (μg/mL)
metal ion release (mg/L)	cell viability (%)
impurity content (%)	toxicity
surface chemistry	
oxidative potential	
exposure dosage (μg/mL)	
exposure time (h)	
cell type	

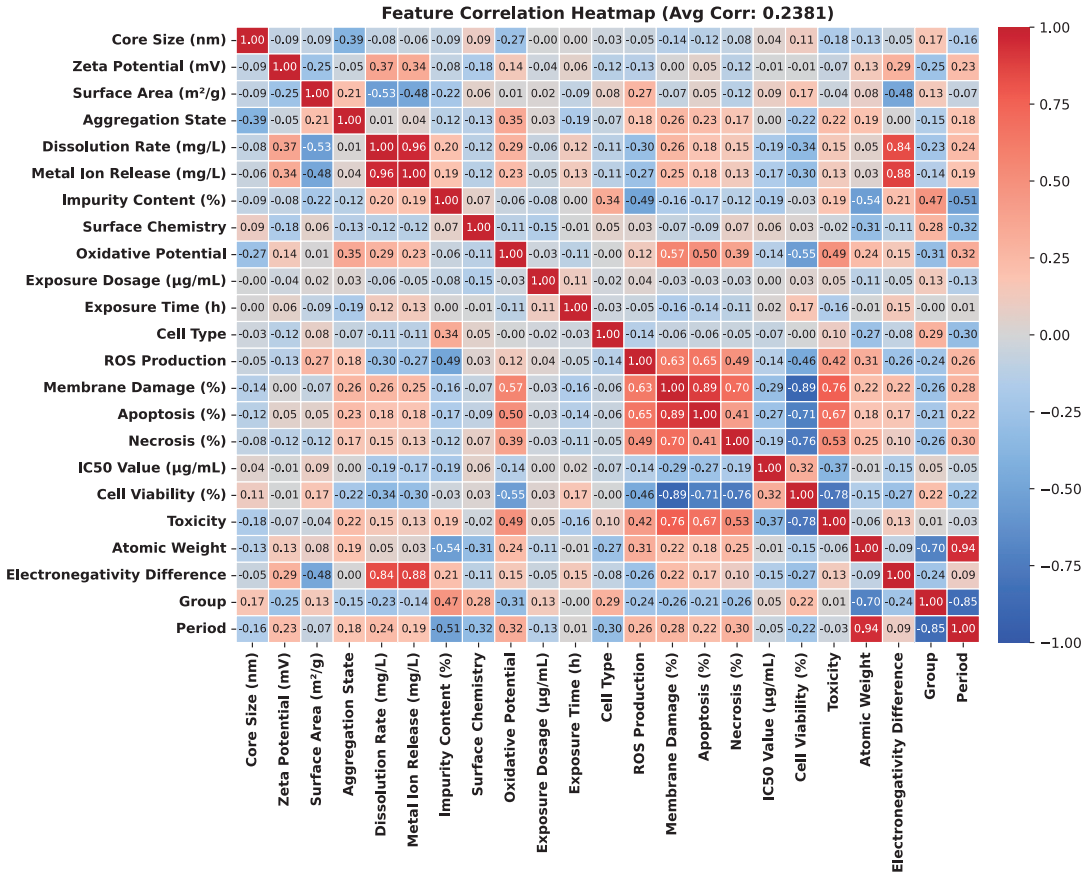


Fig. 1 Correlation matrix for the entire dataset.

weak or no correlation. From the heatmap (**Fig. 1**), it is evident that most feature correlations are close to zero, indicating the absence of strong positive or negative relationships between features. The average Pearson correlation coefficient value of 0.19 further supports the relationships between features, confirming that our dataset is largely uncorrelated. This indicates that our dataset does not suffer from multicollinearity, making it well-suited for feature selection.

In our analysis, we identified that several biological response features including ROS production, membrane damage (%), apoptosis (%), necrosis (%), IC50 value (µg/mL), and cell viability (%) demonstrated significant correlations with the physical properties of the nanoparticles such as core size (nm), surface area (m^2/g), and dissolution rate (mg/L). For instance, ROS production shows a strong correlation with the surface chemistry ($R = 0.69$) and oxidative potential ($R = 0.63$) (as seen in **Fig. 1**), indicating that changes in the nanoparticle surface characteristics are closely linked with the oxidative stress responses in cells. Here, R refers to the Pearson correlation coefficient. Similarly, membrane damage due to apoptosis (0.89) and necrosis (0.70) were highly correlated, reflecting the dependency of these toxicity indicators on each other. These features, being the derivative responses of primary physical interactions, may not provide independent information beyond what is already predictable from the physical features. Furthermore, the zeta potential (mV), which affects the nanoparticle stability and cellular interaction, also shows correlations with several cytotoxic response variables, suggesting that its effects may be implicitly modeled through other physical features. The exclusion of these correlated biological response variables is expected to simplify the model without sacrificing predictive accuracy. This reduction in features not only enhances the computational efficiency but also directs the model's attention toward the most impactful and independent features. In summary, our feature selection strategy prioritizes physical properties that directly influence cytotoxic outcomes, thereby ensuring that our predictive model remains robust.

Our dataset's low correlation ensures that each physical feature contributes uniquely to the prediction of nanoparticle toxicity, preventing misleading results driven by multicollinearity. This independence strengthens feature selection methods like Principal Component Analysis (PCA), allowing us to refine our dataset further without losing critical information.

3. Modeling and preliminary toxicity assessment

3.1 Model selection and training

To evaluate the predictive performance of our

pre-processed dataset, we selected two machine learning models: a gradient boosting machine (GBM) and a support vector machine (SVM) with a Radial Basis Function (RBF) kernel. These models were chosen based on their ability to handle nonlinear relationships, feature dependencies, and structured datasets, which are critical in nanoparticle toxicity prediction. GBM is an ensemble learning method that sequentially improves weak decision trees by minimizing prediction errors (Natekin and Knoll, 2013). This method is particularly useful for datasets where multiple physicochemical and molecular descriptors contribute to toxicity through complex interactions. SVM–RBF, on the other hand, is a robust classification algorithm that maps input features into a higher-dimensional space to capture nonlinear relationships (Suthaharan, 2016). Toxicity prediction is not a simple linear problem, and the ability of SVM–RBF to maximize class separation while handling structured, well-defined clusters makes it a strong candidate for this task. Furthermore, SVM–RBF is resistant to overfitting in moderate-sized datasets, making it a robust alternative to tree-based methods. By comparing these models, we can determine whether toxicity outcomes are best predicted through feature interaction-driven learning (GBM) or high-dimensional decision boundary optimization (SVM–RBF). To validate our model choices, we assessed their accuracy, precision, recall, and F1-score. In addition, plotting the receiver operating characteristic (ROC) curves allows us to compare the classification performance (toxic/non-toxic). We used the following features based on our analysis of lowly correlated features (**Fig. 1**) to train our model to make predictions: core size (nm), shape, surface area (m^2/g), aggregation state, dissolution rate (mg/L), metal ion release (mg/L), surface chemistry, reactive oxygen species potential, exposure dosage (μg/mL), exposure time (h), cell type, atomic weight, electronegativity difference, group, and period. The maximum correlation value for these tested features was less than 0.35, excluding the fundamental molecular features. The dataset was divided into 80 % for training and 20 % for testing for both models. After dataset splitting, a standard scaler was applied to prepare our data for machine learning.

3.2 Model evaluation and metrics

To optimize the model performance, the hyperparameters for both the GBM and SVM–RBF models were tuned using a grid search with *k*-fold cross-validation. For our GBM model, we report an accuracy of 0.83, a precision of 0.78, a recall of 0.77, and an F1-score of 0.78 (**Fig. 2**). For our SVM–RBF, we report an accuracy of 0.81, precision of 0.82, recall of 0.81, and F1-score of 0.78. The GBM model achieved a ROC–AUC (Receiver Operating Characteristic–Area Under the Curve) of 0.83, demonstrating robust predictive capabilities. Although SVM's ROC–AUC was 0.79, which was slightly lower, its good recall value also

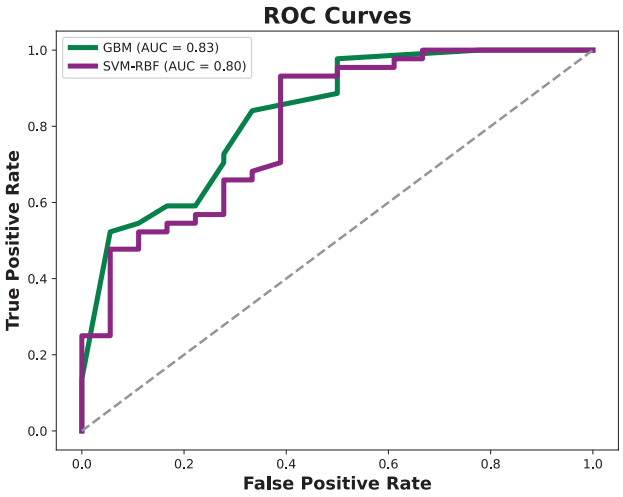


Fig. 2 Comparison of ROC–AUC curves for the selected models to identify the optimal model.

makes it particularly suited for identifying toxic nanoparticles. These results validate our pre-processing approach and feature selection and confirm the dataset's robustness for accurate toxicity prediction using machine learning models. After these initial runs, we applied hyperparameter optimization for the parameters used for GBM and SVM–RBF. GridSearchCV is a technique for finding the optimal parameter values from a given parameter grid. The optimized GBM achieved a train ROC–AUC score of 0.85 and a test ROC–AUC of 0.83, confirming excellent generalization capability. This improvement highlights the GBM's ability to capture complex interactions among the nanoparticle properties effectively. Similarly, optimization improved the SVM model, yielding a train ROC–AUC of 0.80 and a test ROC–AUC of 0.81. Although slightly lower than that of GBM, this score demonstrates SVM's strong classification performance and supports its use as an effective alternative for nanoparticle toxicity prediction.

3.3 Correlation and PCA of the dataset

To ensure the integrity and predictive strength of our model, we performed a two-step analysis involving feature correlation assessment and PCA. Our objective was to select features that are not only relevant to nanoparticle toxicity prediction but also exhibit minimal redundancy. We began by focusing exclusively on the physical, exposure, and fundamental features of the nanoparticles. **Fig. 3** presents the correlation heatmap of the selected features. With an average Pearson correlation coefficient of 0.23, the heatmap reveals that most feature pairs exhibit weak or no linear relationships. This coefficient value indicates a well-balanced dataset with minimal multicollinearity. For instance, features such as core size (nm) and electronegativity difference show little overlap in information content, underscoring their unique contribution to the predictive space. Some moderate correlations, such as between the

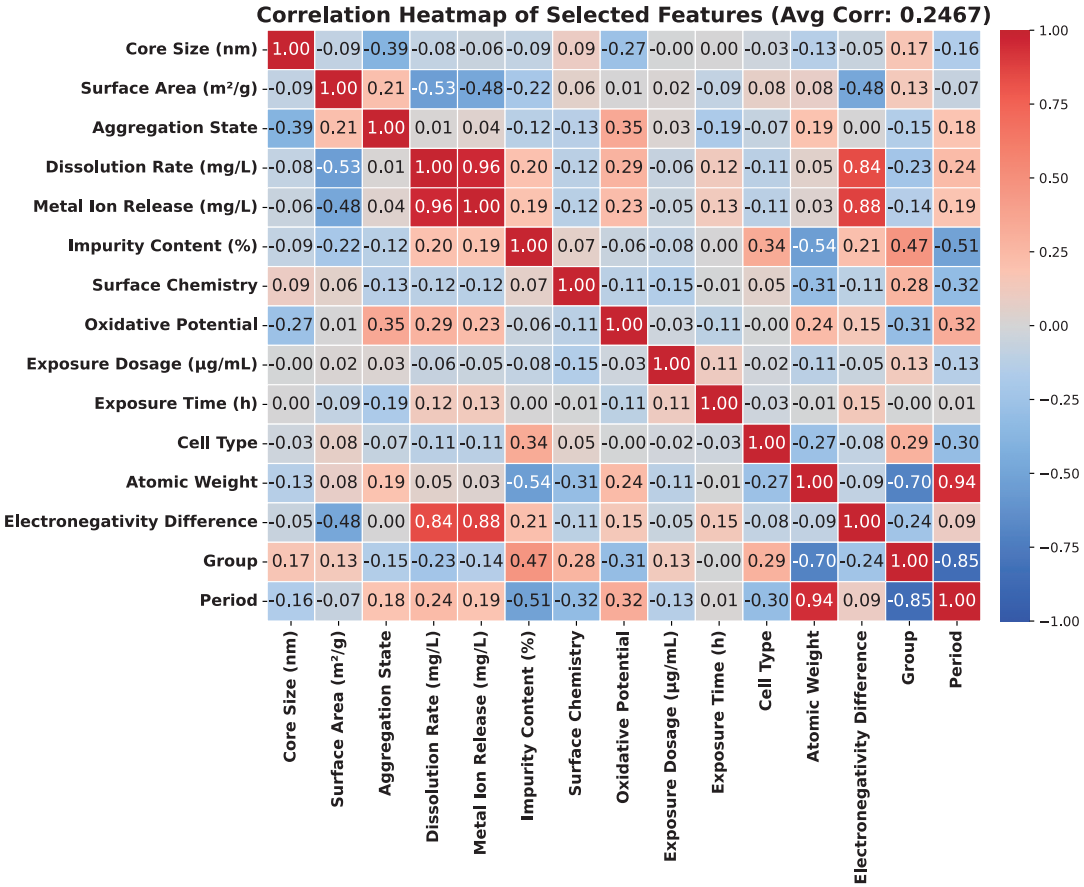


Fig. 3 Correlation matrix for selected features.

dissolution rate and metal ion release, are scientifically expected due to shared physicochemical mechanisms.

To further evaluate feature redundancy and dimensionality, we conducted PCA on the same set of features (**Fig. 4**). The explained variance plot shows that the first five principal components (dissolution rate, aggregation state, core size, surface chemistry, exposure dosage) capture over 70 % of the total variance, and about 10 components are required to explain nearly all the variability in the data. This plot confirms that while the features are largely non-redundant, a few dominant axes of variation exist, indicating meaningful biological and chemical interactions.

We retained features that are both independent and informative, ensuring that our models avoid overfitting (0.85-train, 0.83-test for GBM, 0.86-train, 0.81-test for SVM–RBF) while capturing the essential mechanisms underlying nanoparticle-induced toxicity.

3.4 Confusion matrix analysis

We analyzed the predictions of our model by creating a confusion matrix, as shown in **Fig. 5**. A confusion matrix is a table that summarizes the performance of a classification model by showing the counts of true positives, true negatives, false positives, and false negatives. The confusion matrix on the left is for the GBM, for which we observe a

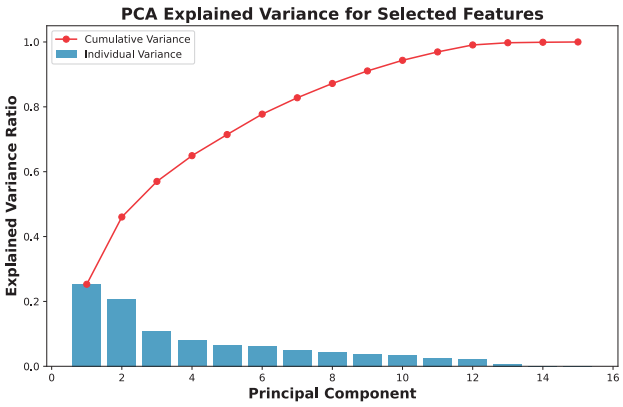


Fig. 4 PCA for explained variance by the number of features.

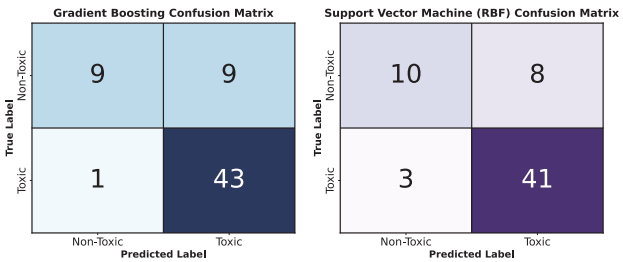


Fig. 5 Confusion matrix comparison for the models showing their general tendencies.

sensitivity of 97.7 % and a specificity of 50 %. Similarly, we obtained a sensitivity of 95.5 % and a specificity of 50 % for the SVM–RBF kernel. The high sensitivity values indicate that the actual positive cases were correctly predicted as positive. These sensitivity values demonstrate that our models rarely missed actual positives, making them well-suited for cytotoxicity prediction where detecting toxic cases is critical. However, a specificity of 50 % for both models implies that only half of the actual negative cases were correctly identified as negative.

3.5 Optimization parameters for toxicity predictions

For the predictions in unseen feature values, we report the trend of exposure dosage versus core size to obtain the toxicity range for our dataset of four metal oxide nanoparticles. From PCA (**Fig. 4**), we trained our data only on five features that exhibited a 70 % variance. For this prediction analysis, we have chosen two features from our test set data set to avoid low interpretability and maintain clarity in visualizing their trends against toxicity. Core Size (nm) and Exposure Dosage (µg/mL) were selected, and other physical features were kept constant, such as Surface Chemistry at 4 which was label encoded for uncoated particle, Oxidative Potential at score 3, etc. Additional constant parameter encoding values can be found in the J-STAGE Data (https://doi.org/10.50931/data.kona.29672717). **Figs. 6** and **7** represent two-dimensional and three-dimensional heatmaps, respectively, of the core size (0–100 nm), exposure dosage (0–1000 µg/mL), and the color scale corresponds to the toxicity probability.

The pockets of non-toxic predictions (blue regions) were observed in specific parameter spaces, particularly within core sizes at 20, 35, 40, 60, and 90 nm, and exposure dosages ranging from 400 to 500 µg/mL (**Fig. 6**). These regions demonstrate that toxicity is not solely dictated by exposure dosage and core size but also by physicochemical properties such as surface chemistry, oxidative potential, and the cell type used. The blue region also indicates the need for searching more data, as there are inconsistent trends due to missing data for some feature ranges. Both visualizations reveal a correlation between the exposure dosage and toxicity probability. As expected, higher nanoparticle concentrations resulted in increased toxicity predictions, aligning with the established cytotoxicity mechanisms. Notably, nearly all configurations are classified as toxic at very high exposure levels (greater than 800 µg/mL). This observation underscores the importance of dosage control in nanomaterial applications, particularly in biomedical contexts where excessive accumulation can lead to significant cellular stress.

Unlike the 2D heatmap, the 3D visualization in **Fig. 7** reveals the gradient and curvature of the response surface, highlighting the nonlinear interactions between the two input features. The surface topology depicts a plateauing behavior at high exposure dosages and small core sizes, where the toxicity probability approaches saturation. We can clearly observe the model capturing the core size of 0–10 nm as toxic for any exposure dosage above a negligible value, as seen in **Fig. 7**. This phenomenon indicates that small core size, leading to a higher surface area, can cause cytotoxicity. Similarly, as the core size increases, the cytotoxicity label tends to reduce, although it always tends to predict toxic for higher exposure dosage.

Despite capturing the expected trends, the model exhibits several limitations. The inconsistency of the three-dimensional surface plot shows uncertainty in certain regions, most likely due to data sparsity in the training dataset. Additionally, the heatmap (**Fig. 7**) contains

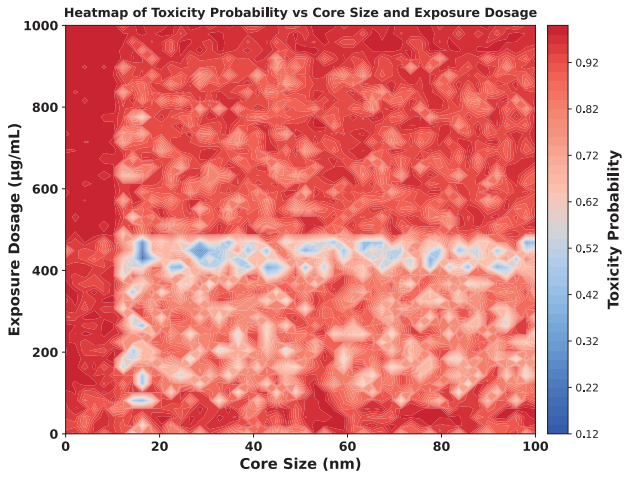


Fig. 6 2D representation of toxicity prediction trends: exposure dosage versus core size.

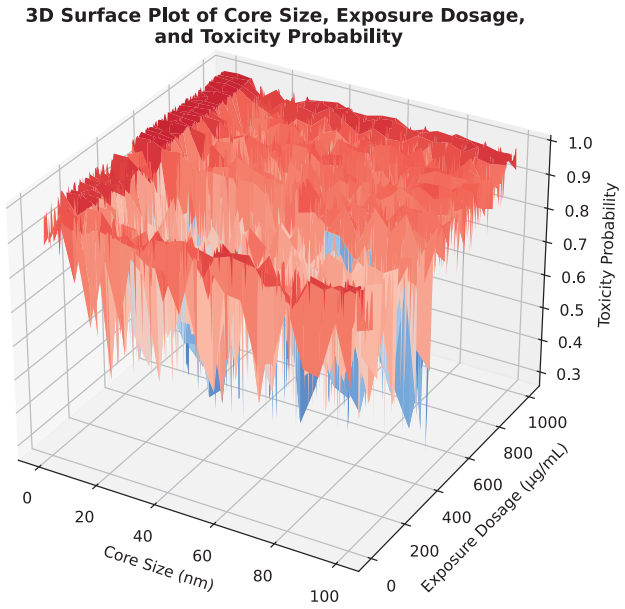


Fig. 7 3D representation of toxicity prediction trends: exposure dosage versus core size.

interpolation artifacts that may not fully reflect the underlying biological trends. These inconsistencies highlight the need for expanded datasets to enhance model generalizability. Furthermore, the predominance of toxic predictions implies a potential class imbalance in the training data, warranting future dataset refinement to ensure adequate representation of non-toxic cases. This dataset refinement can be performed by selecting more non-toxic instances while creating the dataset, using adaptive sampling to maintain class balance. To improve model robustness, work in progress is focusing on dataset balancing techniques such as synthetic augmentation of non-toxic cases and the integration of additional experimental cytotoxicity data. Moreover, validating the identified non-toxic regions through empirical nanoparticle synthesis and biological assays is crucial to confirm the model's predictive reliability. By refining the predictive capabilities and validating against real-world toxicity assessments, this framework can ultimately serve as a guiding tool for safer nanoparticle design.

4. Summary and conclusions

The findings of this study emphasize key takeaways that contribute to advancing machine learning applications in predicting nanotoxicology. First, the importance of comprehensive dataset curation is highlighted, ensuring that the machine learning models are trained on diverse and representative data. In our study, we successfully created a dataset with an average Pearson correlation value of 0.19. Second, we were able to identify five critical parameters (dissolution rate, aggregation state, core size, surface chemistry and exposure dosage) which could explain the 70 % variance in our dataset. Further, when our model was trained on a GBM and SVM–RBF on our dataset, we obtained an accuracy of 83 % and 81 %, respectively, indicating the robustness of our data on models better suited for nonlinear relationship capture. Finally, when our model was used for the prediction of unseen feature values for the five selected features through PCA, we observed three main takeaways. First, the model captures the toxicity of the nanoparticles for all core sizes below 10 nm irrespective of other features. Second, above an exposure dosage value of 800 μg/mL we get toxic ranges irrespective of the core size. Finally, we observed a range of values from 400 to 500 μg/mL where there were pockets of the non-toxic nanoparticle regime. These observations underscore the need to expand the dataset further to capture more continuous and consistent feature trends that have not been captured in our current dataset.

As also experienced during the above reported efforts, a major challenge faced by developers of machine learning models is the lack of reliable, high-quality data that has been replicated identically across experimental investigations reported in the literature. This issue often stems from a limited appreciation and understanding of which particle and powder properties influence their behavior in both dry and wet states and how they do so. Additionally, methodological inconsistencies across studies further complicate data reliability. It is well established that variations in manufacturing processes, environmental conditions during and before measurement, and differences in measurement techniques and instruments can all significantly impact the quality and consistency of data.

A review of the existing literature reveals a broad range of relevant particle and powder properties that must be considered—these include not only basic characteristics but also methodological (e.g., fabrication techniques, measurement tools) and environmental factors (e.g., handling, storage, and biological media used during testing). Compiling and scoring the reliability of source data based on how thoroughly these critical details are documented in publications is desperately needed. This, in turn, would help modelers to assess the significance of their findings relative to the quality of the underlying data in order to develop definitive predictive models.

5. Future work

Despite achieving promising results in toxicity prediction using machine learning, there remain unresolved questions regarding the molecular-scale mechanisms driving cytotoxicity. Predictive models based on bulk physicochemical properties provide correlations but do not fully capture the atomic-level interactions that lead to toxicity outcomes. To address this gap, future work will integrate Molecular Dynamics (MD) simulations and machine learning to establish a multi-scale modeling framework. The next phase of this research will involve MD simulations to model nanoparticle interactions with cellular components such as lipid bilayers and proteins. By simulating aggregation, protein corona formation, and membrane penetration, we aim to provide a mechanistic understanding of how surface chemistry, charge distribution, and nanoparticle dissolution influence cytotoxic effects at the atomic level. These results will be incorporated into our machine learning framework as new features that reflect the dynamic rather than static nanoparticle properties.

Data Availability Statement

We provide access to our dataset, label encodings and procedure for using LLM for data structuring in J-STAGE Data (https://doi.org/10.50931/data.kona.29672717).

Acknowledgments

The authors would like to acknowledge the financial assistance of this work by the Center for Nano-Bio Sensors (CNBS) at the University of Florida.

References

Akter M., Sikder M.T., Rahman M.M., Ullah A.K.M.A., Hossain K.F.B., Banik S., Hosokawa T., Saito T., Kurasaki M., A systematic review on silver nanoparticles-induced cytotoxicity: physicochemical properties and perspectives, Journal of Advanced Research, 9 (2018) 1–16. https://doi.org/10.1016/j.jare.2017.10.008

Alowais S.A., Alghamdi S.S., Alsuhebany N., Alqahtani T., Alshaya A.I., Almohareb S.N., Aldairem A., Alrashed M., Bin Saleh K., Badreldin H.A., Al Yami M.S., Al Harbi S., Albekairy A.M., Revolutionizing healthcare: the role of artificial intelligence in clinical practice, BMC Medical Education, 23 (2023) 689. https://doi.org/10.1186/s12909-023-04698-z

Fröhlich E., The role of surface charge in cellular uptake and cytotoxicity of medical nanoparticles. International Journal of Nanomedicine, 7 (2012) 5577–5591. https://doi.org/10.2147/IJN.S36111

Furxhi I., Murphy F., Mullins M., Arvanitis A., Poland C.A., Practices and trends of machine learning application in nanotoxicology, Nanomaterials, 10 (2020) 116. https://doi.org/10.3390/nano10010116

Krewski D., Daniel A.J., Melvin A., Henry A., Bailar J.C, Kim B., Robert B., Charnley G., Cheung V.G., Green S., et al., Toxicity testing in the 21st century: a vision and a strategy, Journal of Toxicology and Environmental Health, Part B, 13 (2010) 51–138. https://doi.org/10.1080/10937404.2010.483176

Liu R., Jiang W., Walkey C.D., Chan W.C.W., Cohen Y., Prediction of nanoparticles-cell association based on corona proteins and physicochemical properties, Nanoscale, 7 (2015) 9664–9675. https://doi.org/10.1039/C5NR01537E

Manke A., Wang L., Rojanasakul Y., Mechanisms of nanoparticle-induced oxidative stress and toxicity, BioMed Research International, 2013 (2013) 942916. https://doi.org/10.1155/2013/942916

Mishra P.K., Mishra H., Ekielski A., Talegaonkar S., Vaidya B., Zinc oxide nanoparticles: a promising nanomaterial for biomedical applications, Drug Discovery Today, 22 (2017) 1825–1834. https://doi.org/10.1016/j.drudis.2017.08.006

Muralidharan V., Adewale B.A., Huang C.J., Nta M.T., Ademiju P.O., Pathmarajah P., Hang M.K., Adesanya O., Abdullateef R.O., Babatunde A.O., Ajibade A., Onyeka S., Cai Z.R., Daneshjou R., Olatunji T., A scoping review of reporting gaps in FDA-approved AI medical devices, NPJ Digital Medicine, 7 (2024) 273. https://doi.org/10.1038/s41746-024-01270-x

Natekin A., Knoll A., Gradient boosting machines, a tutorial, Frontiers in Neurorobotics, 7 (2013) 21. https://doi.org/10.3389/fnbot.2013.00021

Nel A.E., Mädler L., Velegol D., Xia T., Hoek E.M.V., Somasundaran P., Klaessig F., Castranova V., Thompson M., Understanding biophysicochemical interactions at the nano–bio interface, Nature Materials, 8 (2009) 543–557. https://doi.org/10.1038/nmat2442

Nikolova M.P., Chavali M.S., Metal oxide nanoparticles as biomedical materials, Biomimetics, 5 (2020) 27. https://doi.org/10.3390/biomimetics5020027

Palanivinayagam A., Damaševičius R., Effective handling of missing values in datasets for classification using machine learning methods, Information, 14 (2023) 92. https://doi.org/10.3390/info14020092

Papadiamantis A.G., Afantitis A., Tsoumanis A., Valsami-Jones E., Lynch I., Melagraki G., Computational enrichment of physicochemical data for the development of a ζ-potential read-across predictive model with Isalos Analytics Platform, NanoImpact, 22 (2021) 100308. https://doi.org/10.1016/j.impact.2021.100308

Saptarshi S.R., Duschl A., Lopata A.L., Interaction of nanoparticles with proteins: relation to bio-reactivity of the nanoparticle, Journal of Nanobiotechnology, 11 (2013) 26. https://doi.org/10.1186/1477-3155-11-26

Savage D.T., Hilt J.Z., Dziubla T.D., In vitro methods for assessing nanoparticle toxicity, in: Zhang Q. (Ed.) Nanotoxicity: Methods and Protocols, Springer New York, New York, NY, 2019, pp. 1–29, ISBN: 978-1-4939-8916-4. https://doi.org/10.1007/978-1-4939-8916-4_1

Sun Y., Zhou Y., Rehman M., Wang Y.-F., Guo S., Protein corona of nanoparticles: isolation and analysis, Chem & Bio Engineering, 1 (2024) 757–772. https://doi.org/10.1021/cbe.4c00105

Suthaharan S., Support vector machine, in: Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, Springer US, Boston, MA, 2016, pp. 207–235, ISBN: 978-1-4899-7641-3. https://doi.org/10.1007/978-1-4899-7641-3_9

Vanivska K., Dianová L., Halo Jr M., Štefunková N., Lenický M., Slanina T., Tirpák F., Ivanič P., Stawarz R., Massányi P., Toxicity of nanoparticles on animal and human organism: cell response, Journal of Microbiology, Biotechnology and Food Sciences, 14 (2024) e10844. https://doi.org/10.55251/jmbfs.10844

von Borries K., Holmquist H., Kosnik M., Beckwith K.V., Jolliet O., Goodman J.M., Fantke P., Potential for machine learning to address data gaps in human toxicity and ecotoxicity characterization, Environmental Science & Technology, 57 (2023) 18259–18270. https://doi.org/10.1021/acs.est.3c05300

Xiao X., Trinh T.X., Gerelkhuu Z., Ha E., Yoon T.H., Automated machine learning in nanotoxicity assessment: a comparative study of predictive model performance, Computational and Structural Biotechnology Journal, 25 (2024) 9–19. https://doi.org/10.1016/j.csbj.2024.02.003

Zhang N., Xiong G., Liu Z., Toxicity of metal-based nanoparticles: challenges in the nano era, Frontiers in Bioengineering and Biotechnology, 10 (2022). https://doi.org/10.3389/fbioe.2022.1001572

Zouraris D., Mavrogiorgis A., Tsoumanis A., Saarimäki L.A., del Giudice G., Federico A., Serra A., Greco D., Rouse I., Subbotina J., Lobaskin V., Jagiello K., Ciura K., Judzinska B., Mikolajczyk A., et al., CompSafeNano project: nanoInformatics approaches for safe-by-design nanomaterials, Computational and Structural Biotechnology Journal, 29 (2025) 13–28. https://doi.org/10.1016/j.csbj.2024.12.024

Authors' Short Biographies

Soham Ketan Savarkar is a Graduate Student in the Department of Materials Science and Engineering at the University of Florida. He works in Materials Theory Lab under the guidance of Dr. Richard Hennig. He received his bachelor's in chemical technology from the Institute of Chemical Technology, Mumbai in 2023. His research interest includes first-principles calculations like Density Functional Theory, machine-learning interatomic potentials and computational materials discovery. He has experience with a specialized focus on advancing catalyst development and membrane separation technologies to enhance process efficiency and sustainability.

Authors' Short Biographies

Dr. Jason Gibson is the founder and CEO of Quantum Formatics, a Boston-based startup developing next generation superconducting wires. He received his Ph.D. in Materials Science and Engineering from the University of Florida and bachelor's degree in aerospace engineering from West Virginia University. Dr. Gibson was previously an applied machine learning fellow at Los Alamos National Laboratory. His research focuses on the development and application of AI and first-principles methods for the discovery and design of novel functional materials.

Dr. Vasanthakumar Balasubramanian is a Research Assistant Scientist in the Department of Materials Science and Engineering at the University of Florida. He holds a B.Tech in Chemical Engineering from Bharathidasan University, an M.Tech in Biotechnology from Anna University, and a Ph.D. in Materials Engineering from the Indian Institute of Science (IISc), Bangalore. His multidisciplinary research spans biomaterials, biofilms, mineral separation, nanotechnology, and interfacial engineering. A key focus of his current work involves harnessing nanoparticles and biopolymers to effectively remove microbial contaminants from various substrates. He is also actively engaged in applying artificial intelligence (AI) to predict the structure-property relationships of nanoparticles and assess their potential toxicity, aiming to accelerate the design of safer and more effective nanomaterials.

Prof. Brij M. Moudgil is a Distinguished Professor of Materials Science and Engineering (Emeritus) at the University of Florida. He received his B.E. from the Indian Institute of Science, Bangalore, India and his M.S. and Eng.Sc.D. degrees from Columbia University, New York. His research interests include surfactant and polymer adsorption, dispersion and aggregation of fine particles, adhesion and removal of microbes from surfaces, synthesis of functionalized nanoparticles, antiscaling and surfactant-mediated corrosion inhibitors, photocatalytic degradation of hazardous microbes, and nanotoxicity. He has published more than 400 technical papers and has been awarded over 33 patents. He is a member of the U.S. National Academy of Engineering.

Dr. Richard G. Hennig is the Alumni Professor of Materials Science and Engineering at the University of Florida and Director of the Quantum Theory Project. He received his Diploma in Physics from the University of Göttingen in 1996 and his Ph.D. in Physics from Washington University in St. Louis in 2000. Following postdoctoral and research scientist appointments at the Ohio State University, he joined the faculty of Cornell University in 2006. In 2014, he joined the University of Florida. His research focuses on the development and application of AI and first-principles methods for the computational discovery and design of novel materials, including superconductors, two-dimensional materials, quantum materials, and materials under extreme conditions. He has published over 180 technical papers on the theory and modeling of materials.