# Supporting Information

# Investigating Nanotoxicity: Uncovering Associations and Predictive Factors through Machine Learning Analysis of Published Literature

## S1. Data preprocessing

The two datasets from the previously published articles by Labouta et al. (1) and Gul et al. (2) were merged and used in the present study.

-The study by Labouta et al. (1) contained 2,896 data points (refer to rows): 1. Nanoparticle, 2. Type organic/inorganic, 3. Coat, 4. Diameter, 5. Concentration, 6. Zeta potential, 7. Cells, 8. Cell line (L)/primary cells (P), 9. Human(H) or Animal cells (A), 10. Animal?, 11. Cell morphology, 12. Cell age: Embryonic (E) or adult (A), 13. Exposure time (h), 14. Test, 15. Test indicator, 16. Biochemical metric, 17. Cell viability, 18. Interfernce checked, 19. Colloidal stability checked, 20. Positive control, 21. Publication year, 22. Particle ID, 23. Reference DOI

-Gul et al.'s study (2) contained 4,111 data points: 1. No., 2. Year, 3. Material, 4. Type (Inorganic or organic), 5. Shape, 6. Coat/functional group 7. Synthesis method 8. Surface change, 9. Diameter (nm), 10. Size in water, 11. Size in medium, 12. Zeta in water, 13. Zeta in medium, 14. Cell type 15. No. of cells 16. Human or animal 17. Cell source 18. Cell tissue 19. Cell age 20. Cell line, Primary cells (P, L), 21. Time 22. Concentration 23. Test 24. Test indicator 25. Aspect ratio, 26. Cell viability, 27. PDI, 28. Article ID, 29. DOI

Out of 7,007 data points, we removed specific columns with significant missing values. The columns "shape," "synthesis method," and "charge" were removed due to their high rates of missing data, with over 40% of entries (2,897 out of 7,007 data points) missing. Additionally, the columns "size in water," "size in medium,” “zeta_in_water,” “zeta_in_medium,” “zeta_potential,” and “no_of_cells” exhibited over 45% missing data were removed. The final merged dataset consists of 16 columns: Material, Type, Coat/Functional Group, Diameter (nm), Cell_Type, Human_Animal, Cell_Source, Cell_Tissue, Cell_Morphology, Cell_Age, Cell Line/Primary Cell, Time (hr), Concentration (μg/mL, μM), Test, Test Indicator, and Cell Viability (%).

Next, we addressed the missing values in the 'diameter' column by removing the corresponding rows. Then, the columns “concentrations” from the two datasets were reported in different units: μg/mL and μM. To harmonize the data and reduce sparsity for supervised learning, concentrations in μg/mL were converted to the molar range ($10^{-3}$ to $10^{3}$ μM). This conversion assumed that 100 μg/mL of silver (Ag, MW = 107.87) equals 0.000927 M (or 927 μM). Silver was chosen as the reference material for concentration conversion based on its frequent occurrence in the dataset. Specifically, nanoparticles with coats or functional have significantly larger molecular weights. Thus, the concentrations in μM with the effects of coating were estimated as 0.001 of the corresponding concentration in μg/mL. We used this to factorize the concentration in μg/mL to μM. The ML models were trained based on the concentration range, showing no significant performance changes (data not shown). Notably, the ARM analysis utilized concentration in μg/mL, while the μM concentration was omitted in the dataset.

## S2. Exploratory Data Analysis

The two datasets were combined, leading to a final dataset comprising 7,007 rows. Subsequently, the columns containing many missing values were removed, resulting in the 16 columns as stated in the main manuscript. Then, the missing values of concentration and diameter were removed. Consequently, the final dataset consists of 5,983 rows and 16 columns. Figure S1 was generated using numerical and label-encoded categorical data extracted from the dataset, which consists of 5,983 rows and 16 columns.

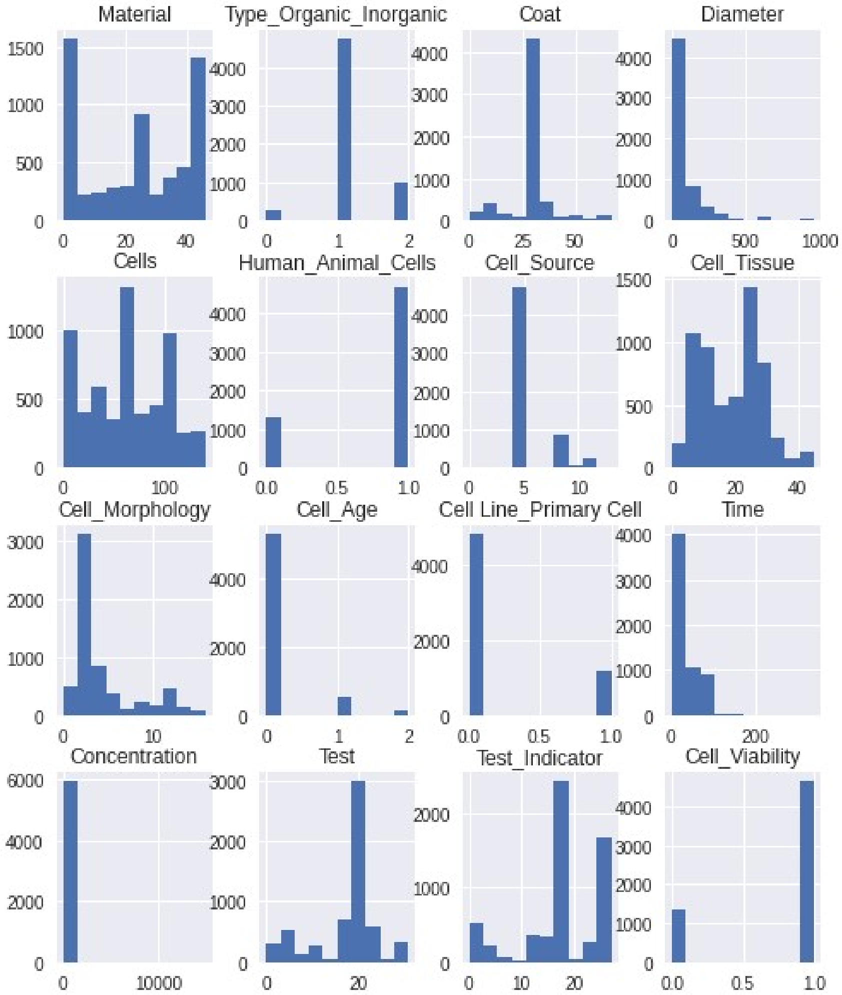


Figure S1: Data distribution after label encoder (Cell viability >= 50%, labeled as 1, and Cell viability < 50% labeled as 0). Only four columns are digits, including

diameter, time, concentration, and cell viability. The remaining 12 columns, including material, type (organic/inorganic), coat, cells, cell source, human/animal cells, cell tissue, cell morphology, cell age, cell line (primary cell), test, and test indicator, have been label-encoded.

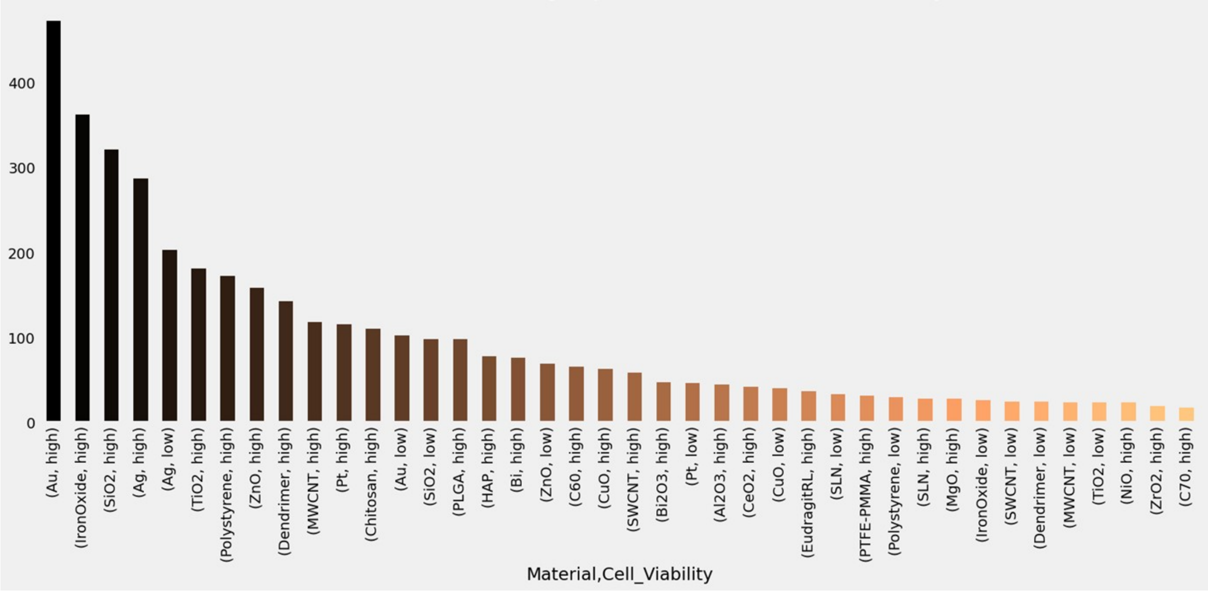


Figure S2: Plot of material type vs. cell viability. The plot displays the dataset's relationship between material type and cell viability. The data reveals that Au (gold nanoparticles) has the highest volume of data points, indicating a high cell viability. In contrast, Ag (silver nanoparticles) is notably associated with increased and lower cell viability.

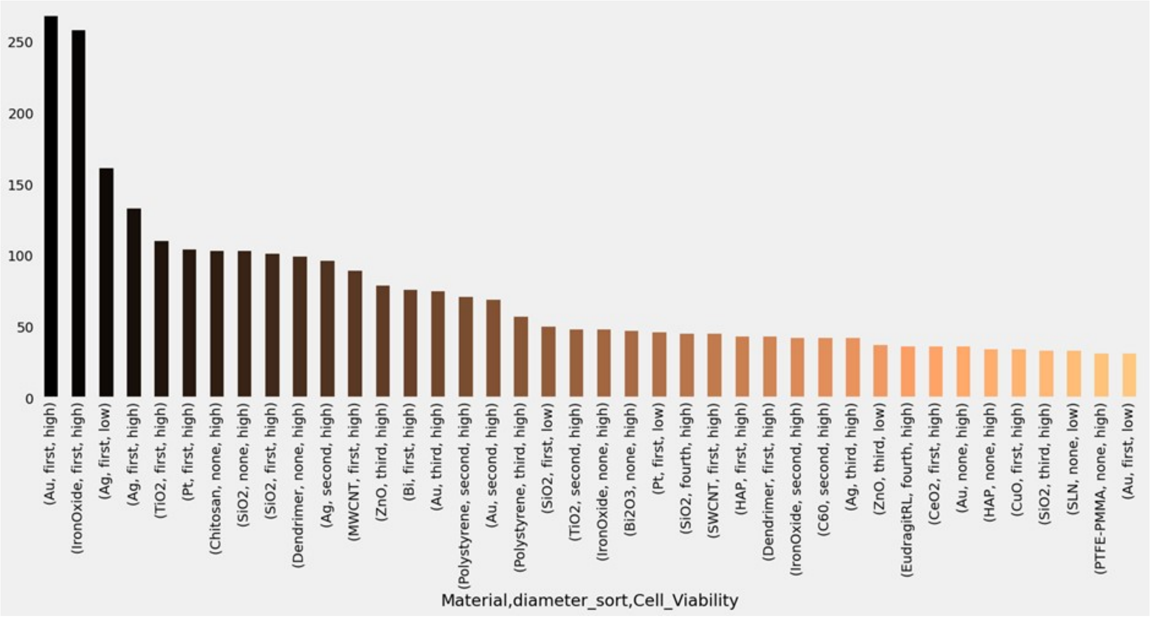


Figure S3: Plot of material, size (diameter), and cell viability. The figure illustrates the distribution of nanoparticles based on their material, size (diameter), and corresponding cell viability. The dataset reveals a substantial presence of small-sized gold (Au) and iron (Fe) nanoparticles with high cell viability. A substantial amount of data on small nanoparticles, such as Ag, Pt, and ZnO, reveals hazardous characteristics with low cell viability.

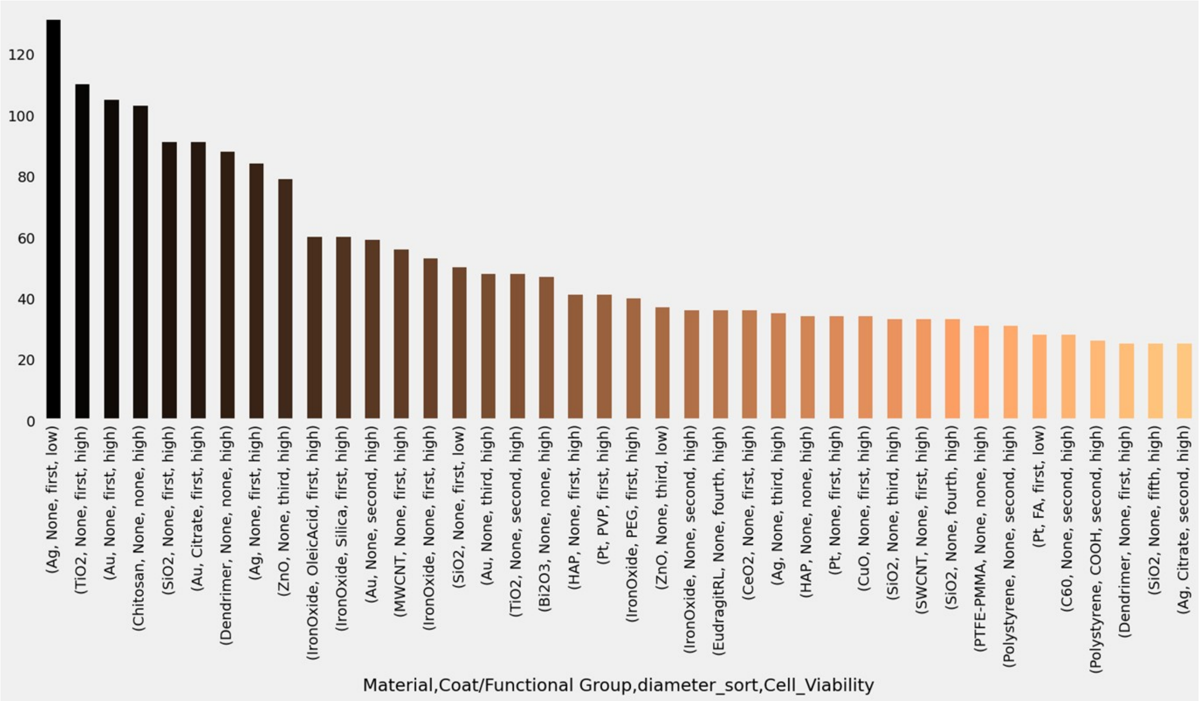


Figure S4: Plot of material, coating/functional group, size, and cell viability. This figure shows the relationship between the material, coating/functional group, size, and cell viability of nanoparticles. The data reveals that small-sized Ag nanoparticles without coating or functional groups exhibit low cell viability, while small $TiO_2$ nanoparticles without coating demonstrate safety. Similar patterns of high cell viability are observed for Au, chitosan, and $SiO_2$ nanoparticles at small sizes.

Ag is quite intriguing. Even though most of the data on Ag demonstrate low cell viability, specific data show good cell viability. We further demonstrated this to see if the viability was impacted by the cell type or test conditions.

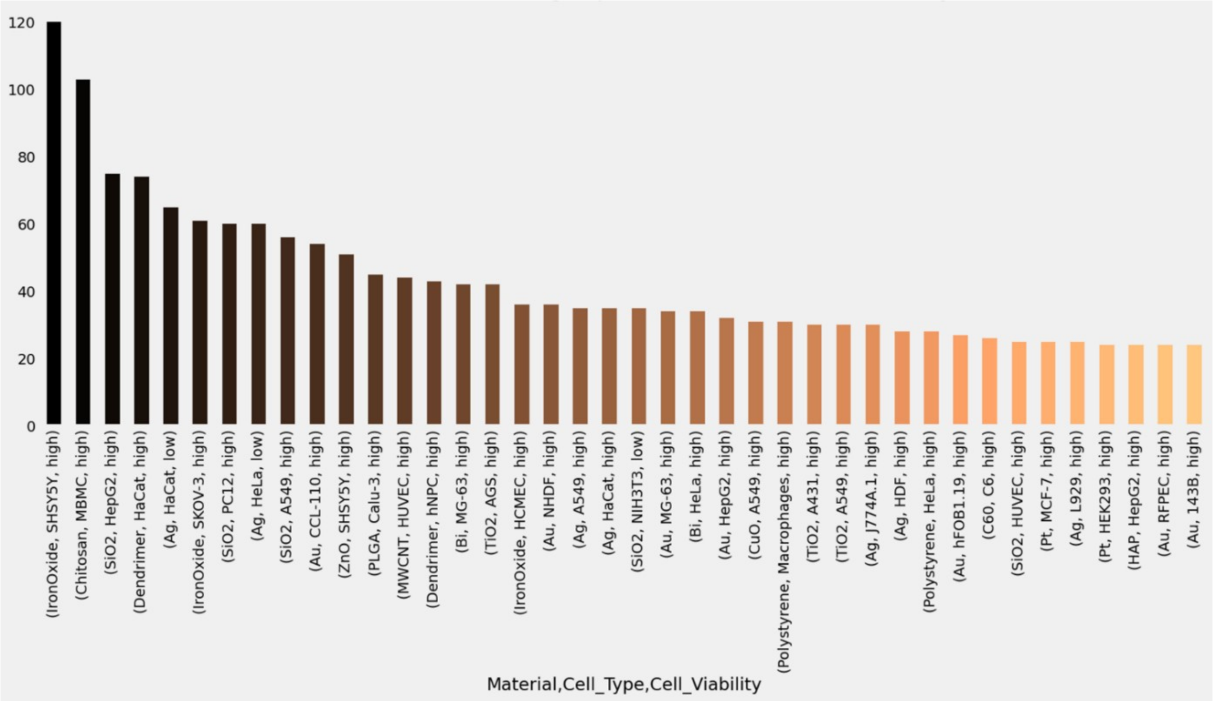


Figure S5: Plot of material, cell type, and cell viability. The plot represents the relationship between material type, cell type, and cell viability of nanoparticles. The data reveals that Ag nanoparticles tested with HeCat and HeLa cell types exhibit low cell viability. There are instances of Ag-tested HeCat cells demonstrating high cell viability. Additionally, Ag nanoparticles tested with A549, J774A1, HDF, and L929 cell types indicate high cell survivability.

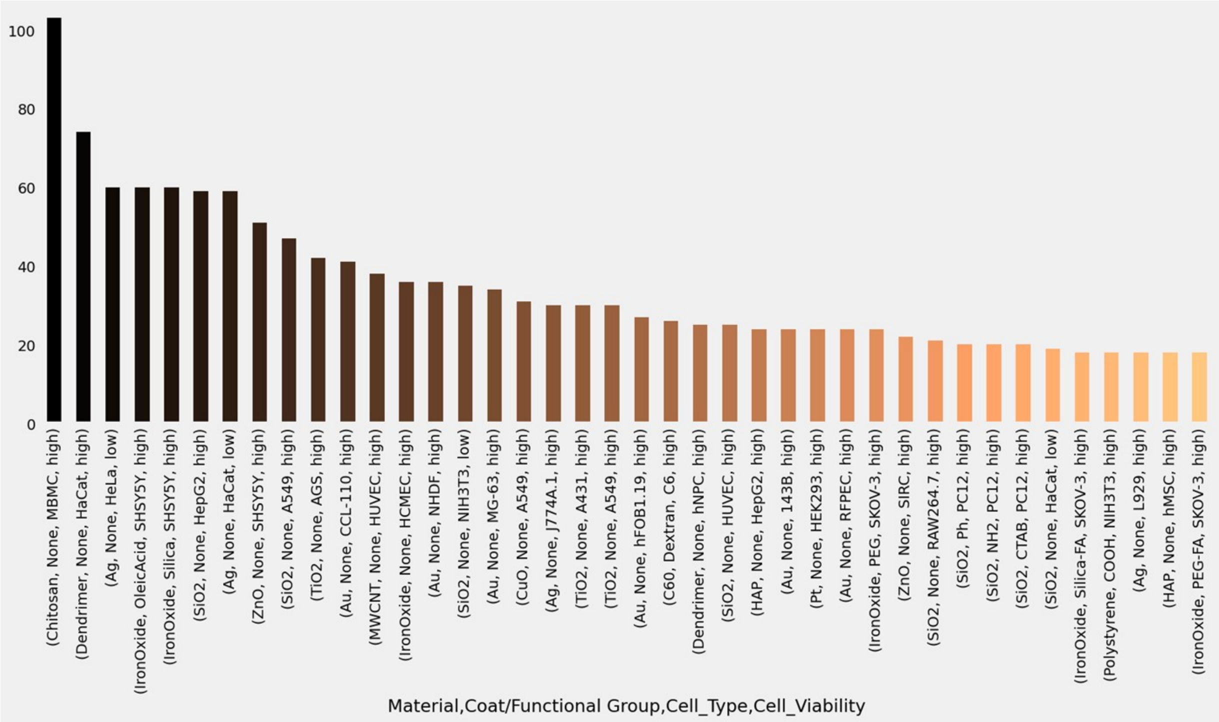


Figure S6: Plot of material, coat/functional group, cell type, and cell viability. Building upon the previous figure's analysis, this figure highlights many non-coated Ag nanoparticles tested with either HeLa or HeCat cell types, demonstrating low cell viability.

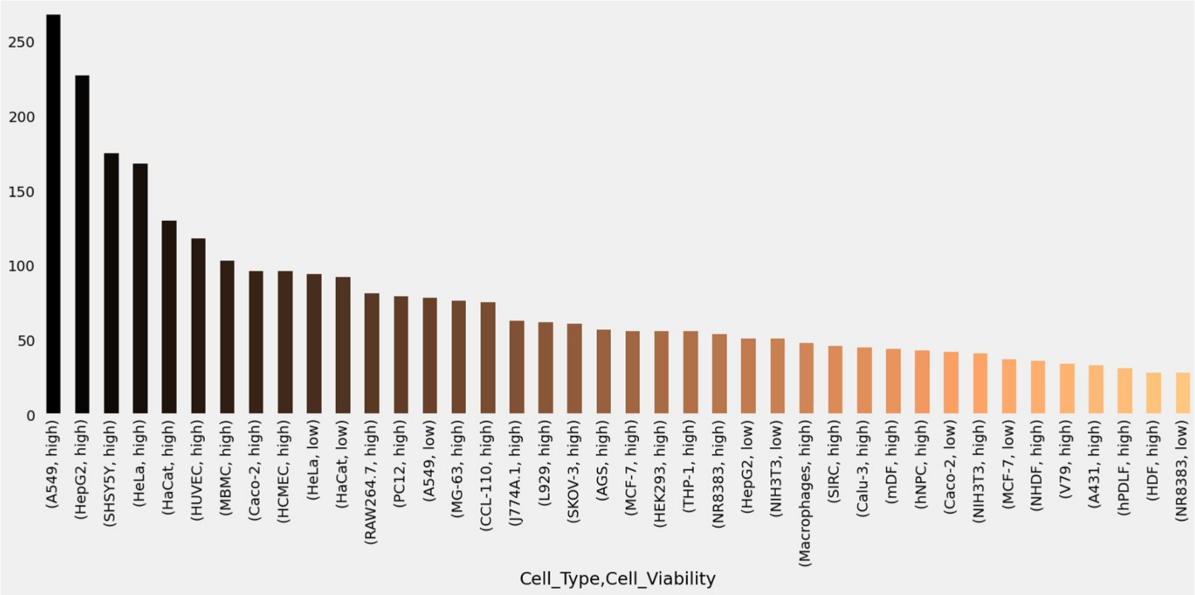
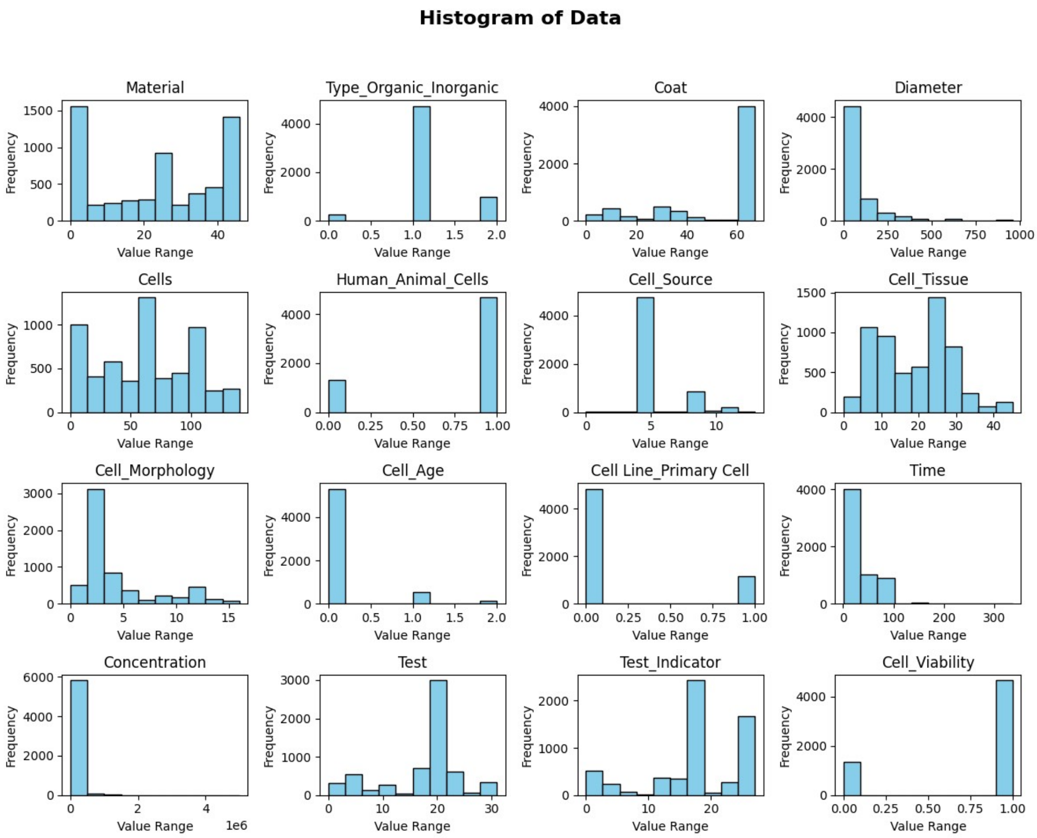


Figure S7: Plot of cell type and cell viability. There are specific data in which lung cancer cell line A549 is associated with a high level of cell viability, and this characteristic is also observed in the hepatocellular carcinoma-derived Hep G2 cell line. Their high cell viability is probably due to cancer cell proliferation. Intriguingly, much of the data reveals associations between the HeLa and HeCat cell types and low cell viability.

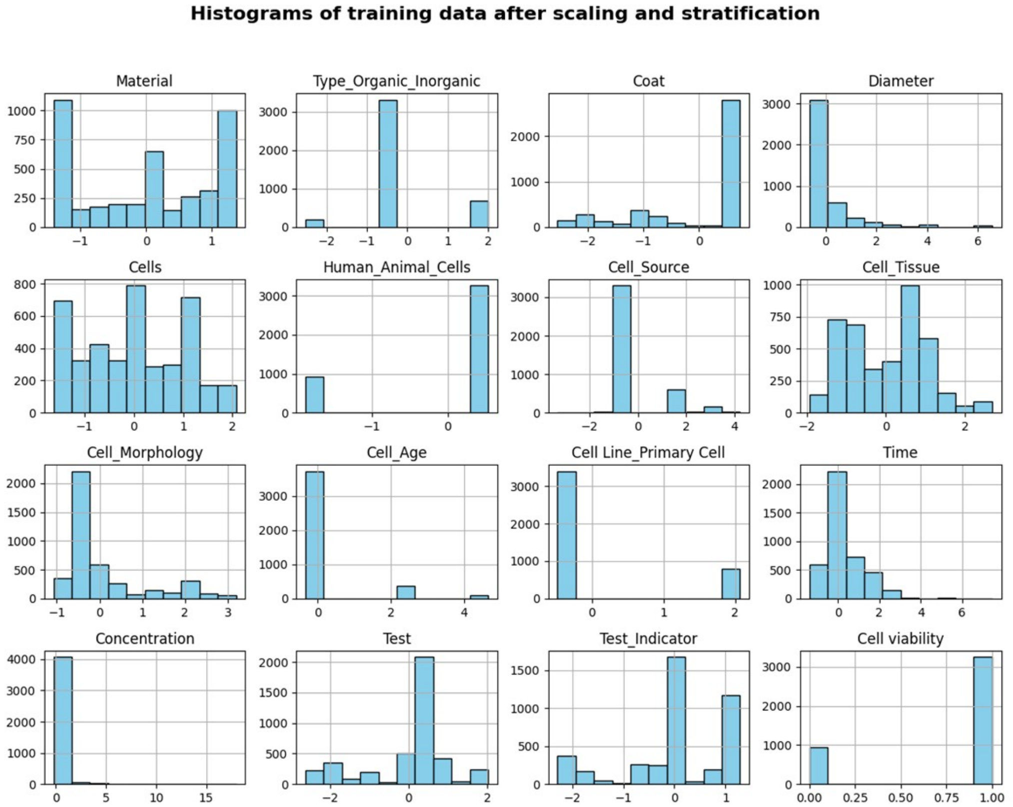
- Stratification

Stratification was used to ensure that the classes are well represented in training and test sets, improving model generalization and reducing bias toward the majority class. Class distribution in the training set (Stratified) is Class 1: 3253, Class 0: 935, and class distribution in the test set (Stratified) is Class 1: 1394 and Class 0: 401. Below are graphs of the preprocessed data. The data was split using stratified sampling to ensure a balanced class distribution. After a split, it was standardized using StandardScaler to prevent potential data leakage, where the model gains prior knowledge of the test set. It standardizes data by subtracting the mean and dividing it by the standard deviation, resulting in a distribution with a mean of 0 and a standard deviation of 1. The mean of each feature is near zero, and the standard deviation is approximately 1.

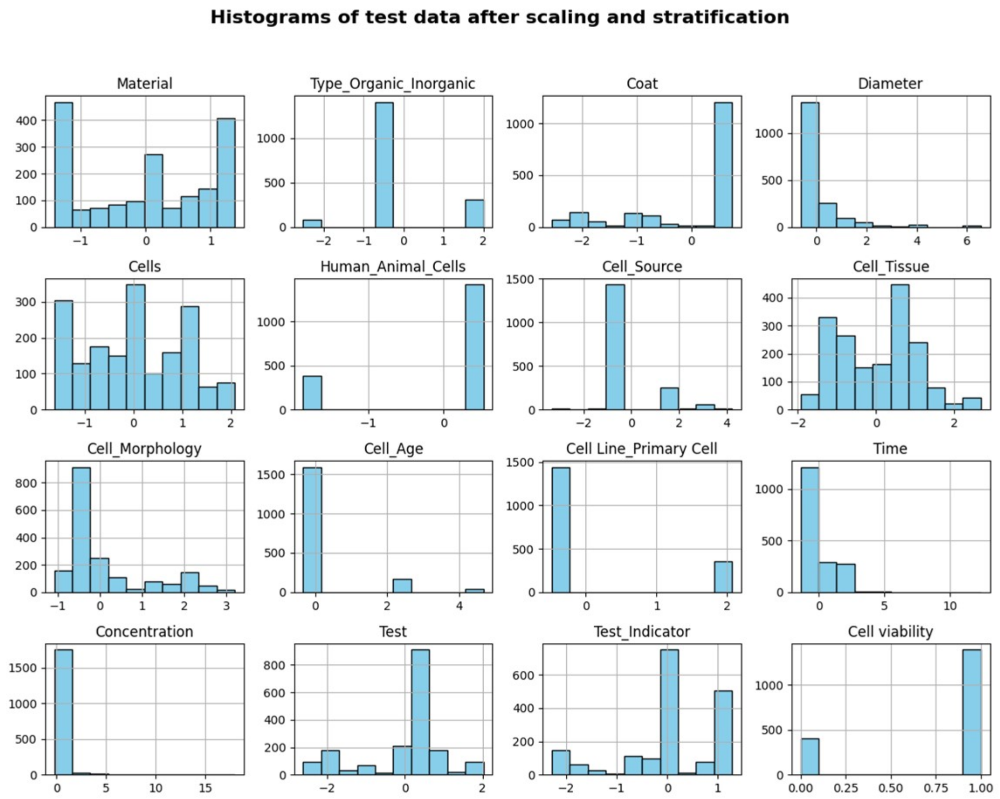
There are also other methods to handle class imbalance, such as cost-sensitive learning, data resampling, etc. For cost-sensitive learning, higher penalties are assigned for misclassifying minority classes, making the model more sensitive to them. However, stratification was selected because it is simple, computationally efficient, and fair in performance evaluation. Methods like cost-sensitive learning and oversampling techniques could be explored for further enhancement.


Histogram of Data
Material
Type_Organic_Inorganic
Coat
Diameter
Cells
Human_Animal_Cells
Cell_Source
Cell_Tissue
Cell_Morphology
Cell_Age
Cell Line_Primary Cell
Time
Concentration
Test
Test_Indicator
Cell_Viability
Frequency
Value Range
1e6
0
20
40
500
1000
1500
0.0
0.5
1.0
1.5
2.0
2000
4000
60
250
750
50
100
0.00
0.25
0.50
0.75
1.00
5
10
30
3000
15
200
300
2
4
6000

(a)


Histograms of training data after scaling and stratification
Material
Type_Organic_Inorganic
Coat
Diameter
Cells
Human_Animal_Cells
Cell_Source
Cell_Tissue
Cell_Morphology
Cell_Age
Cell Line_Primary Cell
Time
Concentration
Test
Test_Indicator
Cell viability
−2
−1
0
1
2
4
6
3
250
500
750
1000
1500
2000
3000
5
10
15
0.00
0.25
0.50
0.75
1.00

(b)



(c)

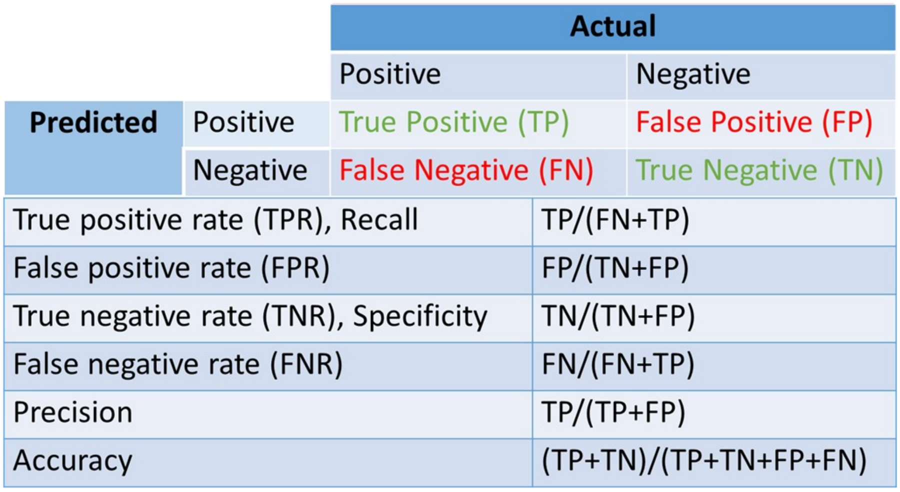
Figure S8: Histogram of data (a) whole data set before scaling and stratification, (b) and (c) the training and test data, respectively, after scaling and stratification.

## S3. Evaluation metrics for classification supervised machine learning

- Receiver Operating Characteristic (ROC)

ROC is a valuable tool for determining the likelihood of a binary result. It plots the false positive rate as the x-axis and the true positive rates as the y-axis for many candidate threshold values between 0.0 and 1.0. The true positive rate (also called sensitivity) can be calculated using the following formula: True positive rate = True positives/(True positives + False negatives). False positive rate = False positives/(False positives + True negatives). The larger values on the y-axis for the positive rates denote a successful prediction, while a small number of false positive rates is expected. A clever model will, on average, place a higher probability on a randomly selected actual positive occurrence than a negative occurrence. Effective models are typically depicted by curves that bow upward and to the top left of the plot. A line drawn diagonally from the plot's bottom left to its top right represents a model with no skill at each threshold, and it has an Area Under Curve (AUC) of 0.5. The AUC describes the area under the ROC curve's integral or a close approximation.

Table S1: Classification evaluation metrics



| | | **Actual** | |
|---|---|---|---|
| | | Positive | Negative |
| **Predicted** | Positive | True Positive (TP) | False Positive (FP) |
| | Negative | False Negative (FN) | True Negative (TN) |

| | |
|---|---|
| True positive rate (TPR), Recall | TP/(FN+TP) |
| False positive rate (FPR) | FP/(TN+FP) |
| True negative rate (TNR), Specificity | TN/(TN+FP) |
| False negative rate (FNR) | FN/(FN+TP) |
| Precision | TP/(TP+FP) |
| Accuracy | (TP+TN)/(TP+TN+FP+FN) |

- Classification report of XGBoost model

The model's performance was also evaluated using precision, recall, and F1-score on the test set. The metrics in Fig. 4(c) show the model's predictive capabilities for class 0 and class 1. The model could perform well in the majority class (class 1, high cell viability) but showed a slight drop in recall for class 0 (low cell viability). When predicting class 1, the model shows a strong predictive capability for all metrics. This suggests that further optimization may be needed to improve the model's performance for the minority class.

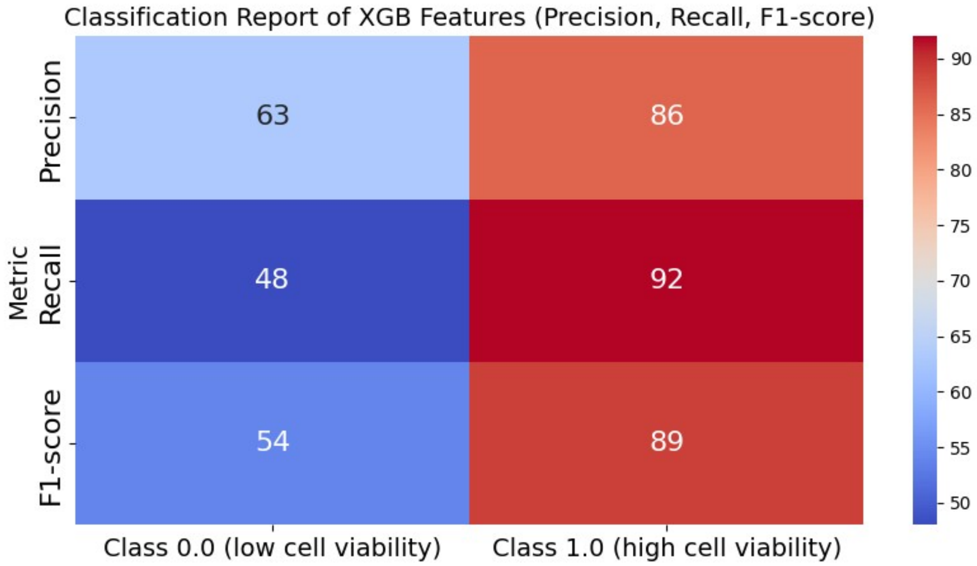


Figure S9: Classification report in % obtained from the 4 features with high feature importance of the XGBoost model.

The model utilizing features selected from ARM demonstrates limited effectiveness in detecting low cell viability (Class 0), capturing recall in merely 35% of actual cases (See Fig. S10). In contrast, it achieves a high recall of 95% for high cell viability (Class 1). Similarly, the F1-score for Class 0 indicates weak performance, highlighting challenges in accurately identifying low cell viability cases, while the model performs well for Class 1 predictions.

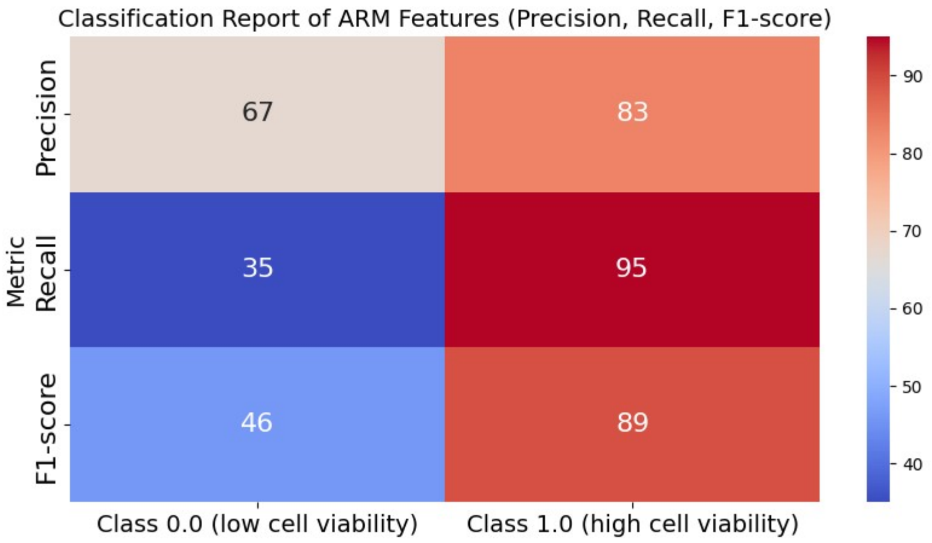


Figure S10: Classification report in % obtained from the four features based on ARM's key features.

The model shows strong predictive capability for high cell viability (Class 1) but struggles with low cell viability (Class 0) predictions. This performance discrepancy is likely due to the imbalanced nature of the dataset, which biases the learning process. Therefore, further improvements such as **data augmentation** or **class rebalancing techniques** are recommended to enhance the model's ability to accurately predict low cell viability cases.

## S4. Unsupervised machine learning: t-SNE algorithm

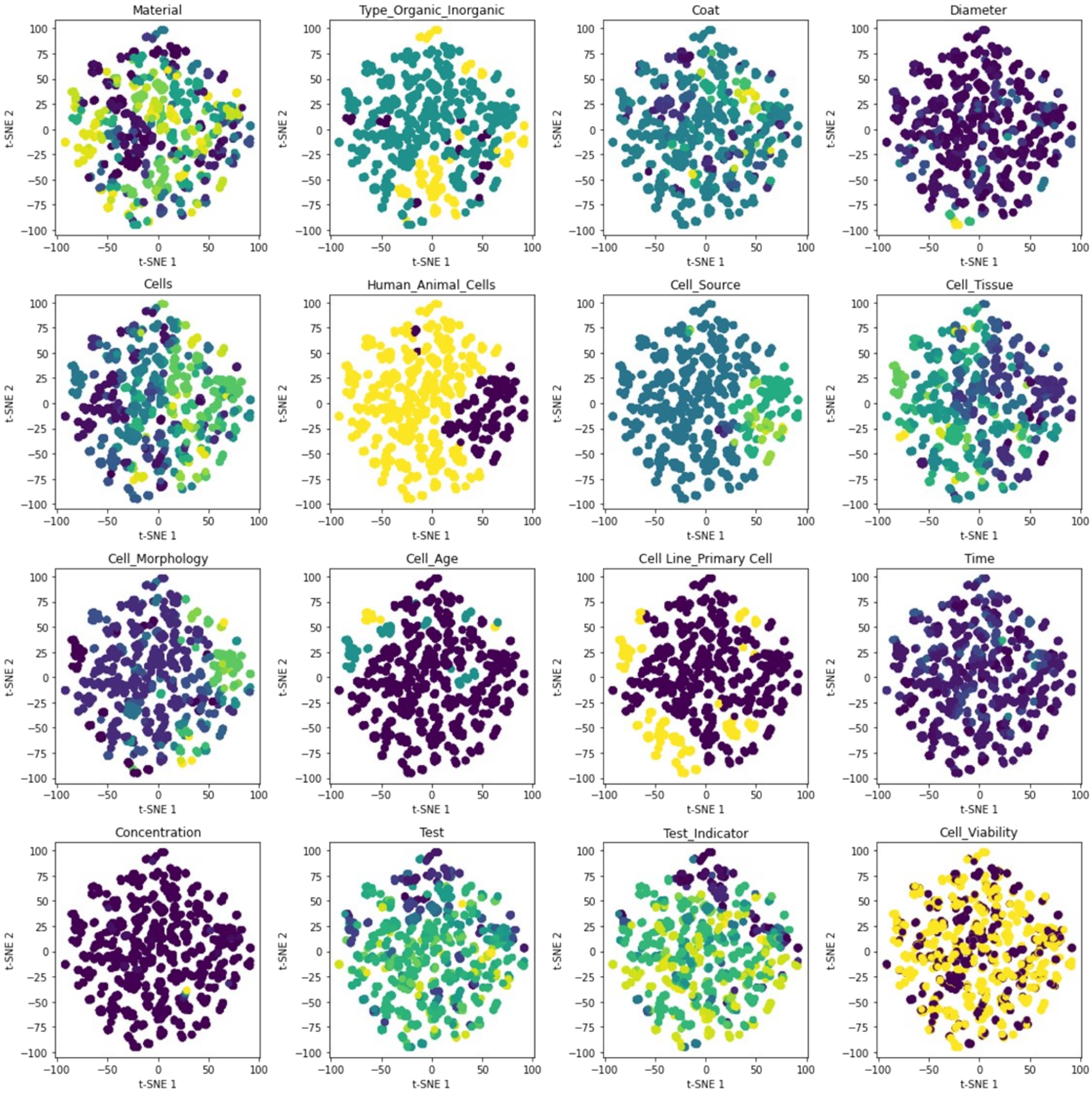


Figure S11: Scatter plots of data analysis using the t-SNE algorithm (n_components = 2, perplexity = 30). The figure caption denotes the scatter plots from data analysis conducted with the t-SNE algorithm.

Table S2: Feature importance based on decision tree (DT) and XGBoost models of the whole dataset (5,983 rows x 16 columns).

| **Feature** | **DT** | **XGBoost** |
|---|---|---|
| Material | 0.337 | 0.094 |
| coat | 0.042 | 0.113 |
| Cell morphology | 0.062 | 0.079 |
| diameter | 0.227 | 0.074 |
| Cell_tissue | 0.025 | 0.068 |
| test | 0.066 | 0.067 |
| Type_organic_inorganic | 0 | 0.06 |
| time | 0.048 | 0.059 |
| Cell_age | 0 | 0.058 |
| Human_animal_cells | 0 | 0.057 |
| Cell_source | 0 | 0.057 |
| Cell line_primary cell | 0.002 | 0.057 |
| Test indicator | 0.043 | 0.056 |
| cells | 0.042 | 0.057 |
| concentration | 0.106 | 0.051 |

Material characters: material, coat, diameter, type of inorganic/organic

Experimental parameters: cell morphology, cell tissue, test, time

## S5: Association Rule Mining (ARM)

The columns (features) to be considered in ARM consist of material, type, shape, coat, synthesis, surface charge, cell_type, test, test-indicator, time_sort human-animal cell, cell_source, cell_tissue, cell_morphology, cell_age, cell_line, time_sort, diameter_sort, and conc_sort. The numeric data were discretized into different ranges as listed below:

| Time Range | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| hr | 1-24 | 25-48 | 49 – 72 | 72 – 96 |

| Diameter Range | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| nm | 1- 30 | 31 – 50 | 50-100 | 101 – 150 | 151-280 | 281-957 |

| Conc Range | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| µg/ml | 0.001 – 0.1 | 0.1 - 1 | 1-3 | 3.1 – 6 | 6.1-10 | 10.1-60 | 60.1 – 150 | 150.1 – 1000 | >1000.1 |

*It should be noted that the concentration unit of µg/ml was chosen in ARM analysis since more data points were listed in this unit. Converting continuous data into categorical features can lead to loss of information during discretization. As a result, ARM findings may oversimplify complex interactions or misrepresent associations. Binning strategies can be employed to mitigate the issues and preserve finer distinctions in the data. The impact of different binning strategies, such as adaptive binning (Automatically adjusting bin widths based on data density), in evaluating ARM metrics should be further explored.

**Table** S3: The example of the materials vs. cell viability by sorting the lift values with confidence greater than 70%.

| **Antecedents** (Material) | **Consequents** (Cell viability) | **support** | **confidence** | **lift** |
|---|---|---|---|---|
| **PLGA** | High | 0.024 | 1.000 | 1.279 |
| **HAP** | High | 0.011 | 0.951 | 1.216 |
| **IronOxide** | High | 0.052 | 0.933 | 1.193 |
| **$TiO_2$** | High | 0.059 | 0.922 | 1.179 |
| **Iron oxide** | High | 0.061 | 0.867 | 1.109 |
| **Bi** | High | 0.022 | 0.864 | 1.104 |
| **Polystyrene** | High | 0.026 | 0.861 | 1.101 |
| **Carbon NP** | High | 0.011 | 0.859 | 1.098 |
| **Au** | High | 0.091 | 0.850 | 1.086 |
| **MWCNT** | High | 0.017 | 0.837 | 1.070 |
| **Carbon Nanotubes** | High | 0.012 | 0.835 | 1.068 |
| **Dendrimer** | High | 0.038 | 0.833 | 1.065 |
| **$SiO_2$** | High | 0.067 | 0.799 | 1.022 |
| **Chitosan** | High | 0.025 | 0.794 | 1.015 |
| **$Al_2O_3$** | High | 0.010 | 0.783 | 1.001 |

Table S4: Association Rule Mining (ARM) Analysis with the two input features (material, coating, and cell viability)

| **Antecedents (coating/functional)** | **consequents** | **support** | **confidence** | **lift** |
|---|---|---|---|---|
| COOH | High | 0.021 | 0.980 | 1.253 |
| Silica | High | 0.012 | 0.955 | 1.221 |
| PEG | High | 0.032 | 0.944 | 1.208 |
| Dextran | High | 0.014 | 0.941 | 1.203 |
| Citrate | High | 0.030 | 0.906 | 1.159 |
| PVP | High | 0.019 | 0.860 | 1.100 |
| PEG-PEI | High | 0.013 | 0.800 | 1.023 |
| Chitosan | High | 0.017 | 0.786 | 1.005 |
| NH2 | High | 0.019 | 0.766 | 0.979 |
| PEI | High | 0.014 | 0.721 | 0.921 |

Table S5: Multiple antecedents (features: material, coating, type of material). The multiple columns relating to the material characteristics, including material, coating, and type of material) and cell viability were used.

| Antecedents | Consequents | support | confidence | lift |
|---|---|---|---|---|
| SLN | low | 0.011 | 0.602 | 2.705 |
| SLN, O | low | 0.011 | 0.602 | 2.705 |
| None, Ag, I | low | 0.047 | 0.535 | 2.404 |
| None, Ag | low | 0.047 | 0.535 | 2.404 |
| Ag | low | 0.059 | 0.469 | 2.107 |
| Ag, I | low | 0.059 | 0.469 | 2.107 |
| I, ZnO | low | 0.022 | 0.355 | 1.597 |
| ZnO | low | 0.022 | 0.355 | 1.597 |
| None, I, ZnO | low | 0.020 | 0.337 | 1.515 |
| None, ZnO | low | 0.020 | 0.337 | 1.515 |
| Pt, I | low | 0.013 | 0.311 | 1.397 |
| Pt | low | 0.013 | 0.311 | 1.397 |
| None, I, $SiO_2$ | low | 0.018 | 0.291 | 1.306 |
| None, $SiO_2$ | low | 0.018 | 0.291 | 1.306 |
| None, I | low | 0.135 | 0.284 | 1.278 |

Table S6: Multiple antecedents (features: material, cell tissue, and test)

| **Antecedents** | **Consequents** | **support** | **confidence** | **lift** |
|---|---|---|---|---|
| Cervix, Ag | low | 0.021 | 0.774 | 3.477 |
| SLN | low | 0.011 | 0.602 | 2.705 |
| MTT, Ag | low | 0.017 | 0.561 | 2.522 |
| Skin, Ag | low | 0.024 | 0.549 | 2.468 |
| Ag | low | 0.059 | 0.469 | 2.107 |
| Embryo | low | 0.014 | 0.446 | 2.003 |
| Cervix | low | 0.030 | 0.399 | 1.794 |
| ZnO | low | 0.022 | 0.355 | 1.597 |
| AlamarBlue | low | 0.015 | 0.329 | 1.477 |
| Pt | low | 0.013 | 0.311 | 1.397 |
| NR | low | 0.011 | 0.302 | 1.357 |
| Skin | low | 0.036 | 0.299 | 1.343 |

Table S7: ARM analysis between experimental parameters and cell viability: considering the lift more than 1.2. Low means "Low cell viability".

| Antecedents | Consequents | support | confidence | lift |
|---|---|---|---|---|
| **Cell morphology and viability** | | | | |
| Keratinocyte | low | 0.022 | 0.432 | 1.982 |
| low | Keratinocyte | 0.022 | 0.100 | 1.982 |
| **Test and viability** | | | | |
| low | LDH | 0.016 | 0.074 | 1.949 |
| LDH | low | 0.016 | 0.425 | 1.949 |
| AlamarBlue | low | 0.013 | 0.329 | 1.508 |
| low | AlamarBlue | 0.013 | 0.060 | 1.508 |
| NR | low | 0.011 | 0.271 | 1.242 |
| low | NR | 0.011 | 0.049 | 1.242 |
| **Test indicator and viability** | | | | |
| LDH activity assay kit | low | 0.016 | 0.400 | 1.835 |
| low | LDH activity assay kit | 0.016 | 0.072 | 1.835 |
| AlamarBlue | low | 0.016 | 0.350 | 1.605 |
| low | AlamarBlue | 0.016 | 0.075 | 1.605 |
| low | toluene red | 0.011 | 0.049 | 1.212 |
| toluene red | low | 0.011 | 0.264 | 1.212 |

Table S8: ARM of the features from experimental conditions (concentration, time, cell tissue, and test). LDH is the test; one denotes the time_sort in the range of 24 hrs., and conc_sort has none (no data reported). It is seen that cell tissue is an embryo, indicating the high confidence of the low viability.

| Antecedents | Consequents | Support | Confidence | Lift |
|---|---|---|---|---|
| (none, one, LDH) | low | 0.013988 | 0.457944 | 2.101084 |
| (Embryo) | low | 0.011704 | 0.445652 | 2.044688 |
| (none, LDH) | low | 0.015701 | 0.44 | 2.018756 |
| (one, LDH) | low | 0.014416 | 0.43913 | 2.014766 |
| (LDH) | low | 0.016129 | 0.424812 | 1.949072 |

Table S9: ARM analysis of the antecedents' features: material, type of organic or inorganic, coat, and diameter.

| Antecedents | Consequents | support | confidence | lift |
|---|---|---|---|---|
| (Ag, None, first, I) | (low) | 0.03811 | 0.646489 | 2.966145 |
| (Ag, None, first) | (low) | 0.03811 | 0.646489 | 2.966145 |
| (Ag, first, I) | (low) | 0.042963 | 0.59252 | 2.718528 |
| (Ag, first) | (low) | 0.042963 | 0.59252 | 2.718528 |
| (SLN, O) | (low) | 0.010277 | 0.590164 | 2.70772 |

Ag is material, and the type of organic/inorganic is I. First is the diameter_ranked, referring to small sizes; none is from "Coat." Ag, CuO (data not shown), ZnO (data not shown), Pt (data not shown), and $SiO_2$ (data not shown) show a significant impact on the low cell viability.

## S6. Chi-square test of antecedents and consequents from ARM analysis

The observed frequencies can be obtained from the contingency_table, while the expected frequencies can be computed based on the assumption that antecedents and consequents are independent $E_{ij} = \frac{(row\ total\ x\ column\ total)}{grand\ total}$. The Chi-square test is calculated as: $\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$, where $O_{ij}$ is the observed frequency and $E_{ij}$ is the expected frequency. The p-value indicates the probability of observing such a relationship by chance.

A p-value was then computed to investigate the statistical significance, in which a p-value less than 0.05 indicates a significant association. The function chi_square_test (not shown) applied the Chi-Square test to each rule (antecedent -> consequent) using contingency tables. Then, a bar plot in the figure below visualizes the top 10 significant association rules based on p-values in agreement with the ARM analysis.

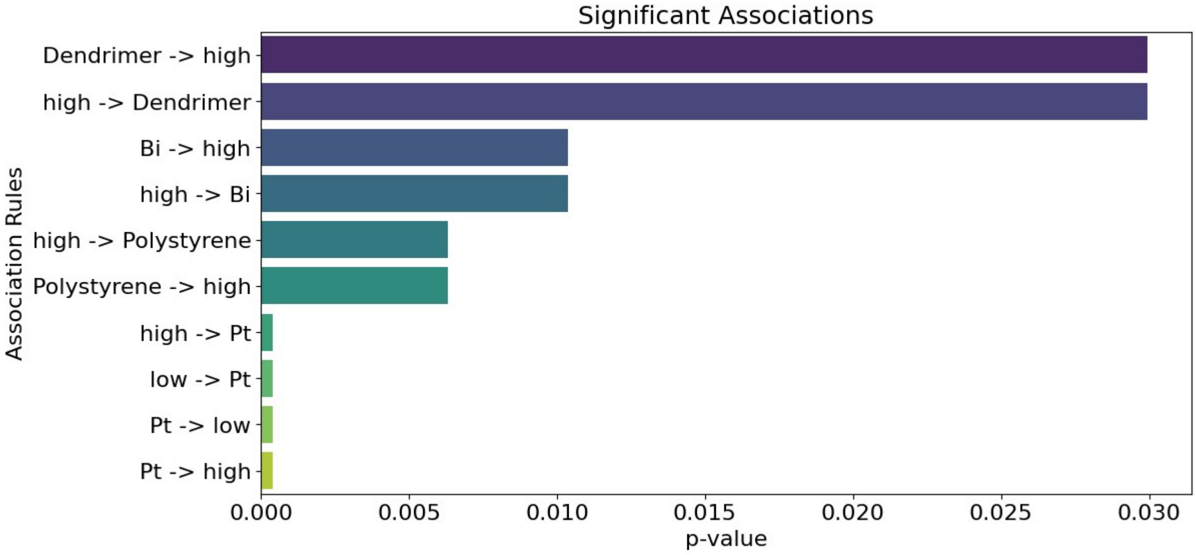


Figure S12: Chi-square test of antecedents and consequents from ARM analysis.

## References

1. Labouta HI, Asgarian N, Rinker K, Cramb DT. Meta-Analysis of Nanoparticle Cytotoxicity via Data-Mining the Literature. ACS Nano. 2019;13(2):1583-94.

2. Gul G, Yildirim R, Ileri-Ercan N. Cytotoxicity analysis of nanoparticles by association rule mining. Environmental Science: Nano. 2021;8(4):937-49.