NANO · MICRO small
www.small-journal.com

# Quantitative Prediction of Inorganic Nanomaterial Cellular Toxicity via Machine Learning

*Nikolai Shirokii, Yevgeniya Din, Ilya Petrov, Yurii Seregin, Sofia Sirotenko, Julia Razlivina, Nikita Serov,* and Vladimir Vinogradov**

**Organic chemistry has seen colossal progress due to machine learning (ML). However, the translation of artificial intelligence (AI) into materials science is challenging, where biological behavior prediction becomes even more complicated. Nanotoxicity is a critical parameter that describes their interaction with the living organisms screened in every bio-related research. To prevent excessive experiments, such properties have to be pre-evaluated. Several existing ML models partially fulfill the gap by predicting whether a nanomaterial is toxic or not. Yet, this binary categorization neglects the concentration dependencies crucial for experimental scientists. Here, an ML-based approach is proposed to the quantitative prediction of inorganic nanomaterial cytotoxicity achieving the precision expressed by 10-fold cross-validation (CV) $Q^2 = 0.86$ with the root mean squared error (RMSE) of 12.2% obtained by the correlation-based feature selection and grid search-based model hyperparameters optimization. To provide further model flexibility, quantitative atom property-based nanomaterial descriptors are introduced allowing the model to extrapolate on unseen samples. Feature importance is calculated to find an interpretable model with optimal decision-making. These findings allow experimental scientists to perform primary in silico candidate screening and minimize the number of excessive, labor-intensive experiments enabling the rapid development of nanomaterials for medicinal purposes.**

## 1. Introduction

Machine learning (ML) has already changed the field of organic chemistry by making molecular physicochemical properties prediction possible including highest occupied molecular orbital/lowest unoccupied molecular orbital energies,[1,2] logP,[3] solubility,[4] etc. Therefore, ML streamlines the targeted discovery of molecular structures with predefined properties. Retrosynthesis planning also becomes less and less challenging with the appearance and development of new approaches, namely i) sequence-to-sequence models[5] working with simplified molecular input line entry system (SMILES) molecule representations and treating the problem as translational, ii) graph neural networks (GNNs),[6] especially graph convolutional networks (GCNs),[7] and iii) hybrid algorithms. Even property-based molecule generation – which is classified as a much harder inverse problem – is possible with variational autoencoders (VAEs)[8] and generative adversarial networks (GANs),[9] though multi-property molecule generation still remains challenging and suffers from relatively low precision.[10] For instance, a successful de novo design of novel antibiotics[11] was demonstrated, where eight structurally unique active molecules were generated distinct from those presented in the initial dataset. Moreover, there are numerous models predicting organic molecule cytotoxicity.[12–14]

Molecular descriptors include bags of bonds, SMILES representations, weighted graphs, electron density maps, Coulomb matrices, etc.[15] and allow building accurate predictive models. At the same time, nanomaterials represent much more complex structures, which are harder to describe in a quantitative manner. The deficiency of material descriptors, as well as experimental data, limits the application of artificial intelligence (AI) to materials science. Despite these challenges, AI-aided materials science has shown remarkable success in the prediction of such material properties as semiconductor band gaps,[16] metamaterial optical responses,[17] material crystalline density,[18] space groups[19] and even some nanomedicine-related properties.[20–23] This became possible due to the descriptors of bulk nanostructured materials, for example topological[24] and crystal-based,[25,26] to name a few. However, when it comes to nanoparticles, these descriptors become insufficient, as they bear no information about their structure and surficial properties. Accurate representation of nanomaterials requires various

N. Shirokii, Y. Din, I. Petrov, Y. Seregin, S. Sirotenko, J. Razlivina, V. Vinogradov
International Institute "Solution Chemistry of Advanced Materials and Technologies"
ITMO University
191002 Saint-Petersburg, Russian Federation
E-mail: vinogradov@scamt-itmo.ru
N. Serov
Advanced Engineering School
Almetyevsk State Oil Institute
Almetyevsk, Russia
E-mail: serov@scamt-itmo.ru
The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/smll.202207106.

**DOI: 10.1002/smll.202207106**

 

descriptors based on nanomaterial structure,[27] atom- and atom distance-based parameters,[28] quantum mechanics calculations,[29] SMILES representations,[30] periodic table constants,[31] physicochemical properties such as size, shape, and surface charge,[32] and experimental data.[33] The descriptors are used to predict various bio-properties as protein corona formation,[34] cellular uptake,[35] ecotoxic effects[36] as well as cytotoxicity,;[37,38] the latter represents a crucial parameter screened in almost any nanomaterial-related research. At the same time, this parameter strongly depends not only on nanomaterial chemical composition and surface modifications but also on shape, surface area, linear sizes, size distribution, surface charge, porosity, etc.[38–42] making it hard to presume heuristically. Each therapeutic or/and diagnostic system must pass this test, and it seems reasonable to have a model able to pre-evaluate this parameter a priori, to circumvent the synthesis of nanomaterials to be thrown in the nearest future. However, current models predict cytotoxicity only in binary toxic/non-toxic classification, which gives little information about concentration dependencies for experimental scientists.[43,44]

In this paper, we propose an ML-based approach towards the quantitative in vitro cytotoxicity prediction on inorganic nanomaterials, where the final LGBM Regressor model achieved 10-fold cross-validation (CV) $Q^2$ of 0.86 and RMSE of 12.2%. Correlation-based feature selection and hyperparameter grid search allowed to screen a large model space of 40 ML models and further optimize the best model. Atom-based quantitative nanomaterial descriptors were introduced to provide model flexibility, namely, the ability to work beyond the chemical entities presented in the dataset. Feature importance was used to find an interpretable model with an optimal decision-making process. The model promises to perform the primary and fully computational candidate toxicity screening, excluding excessive, labor-intensive experiments and enabling rapid development of nanomaterials for biomedical purposes.

## 2. Results

Cytotoxicity of nanomaterials represents a complex process, where such parameters as surface charge, surface coating, chemical composition, hydrodynamic diameter, surface area, nanoparticle concentration, and cellular descriptors determining the nanoparticle environment play an imperative and synergistic role. Thus, building a reliable and interpretable ML model requires an extensive data collection that takes all parameters into consideration. Moreover, because of the sparsity and scarcity of materials science-related data, even a larger dataset is needed to provide accurate ML model training. For this purpose, all available datasets on nanomaterial cellular toxicity were aggregated[39,45] and then supplemented with the manually collected data from research articles. It resulted in an overall number of 8076 raw samples, predominantly consisting of metals and oxides (Figure S1, Supporting Information). Statistical analysis of categorical and numeric parameters proved no significant categorical imbalance for this data (Figure S2 and S3; Tables S1 and S2, Supporting Information). To check for interdependent and non-overlapping parameters, the correlation matrix was calculated on the data before further manipulations. As a result, it was shown that categorical features should be properly encoded before any ML model development (Figure S4, Supporting Information). Since manually collected data always contains outliers and mistakes, it was carefully cleaned using statistical methods and filtering (see Methods section in Supporting Information). It is more complicated with data sparsity since even after manual data collection blank values comprise substantial part of almost every dataset comprise a substantial part of almost every dataset parameter (**Figure 1**a). Deleting all such samples reduces dataset size drastically. Therefore, proper filling methods should be implemented. Parameters including electronegativity, ionic radius, and molecular weight are crucial in terms of chemical composition, so they were additionally mined using Python libraries for chemistry. Being an extremely important parameter, which characterizes nanoparticle surrounding environment crucial to describe the toxicity phenomenon, zeta-potential was filled for this task using its dependence on the calculated mean charge density (mCD) descriptor introduced in our previous work (Figure S5, Supporting Information).[46] To reduce the number of correlations within the data, feature selection was implemented resulting in low level of parameter intercorrelations (Figure 1b). To simplify the categorical values processing by ML model, they were encoded as conditional probabilities of label X given the viability range Y to be predicted (see Methods section in Supporting Information). Due to high data sparsity and the fact that too massive data filling can lead to an incorrectly trained ML model, the processed dataset contained 3087 samples after all the operations (see Supplementary file 1 or GitHub repository), which is enough for the development of high-quality ML models. Clustering by t-distributed stochastic neighbor embedding (t-SNE) showed that the collected dataset allows to successfully differentiate between the chemical compositions, coated and naked nanoparticles, as well as viability values (Video S1, Supporting Information). Most of the compounds form separate groups, which are important for ML models to learn meaningful patterns from the data (**Figure 2**a; see Methods section in Supporting Information). As a result, in this stage, all data was properly curated including data merge, outliers and blank values processing, correlation-based feature selection and synthesis, categorical features encoding, as well as data analysis.

To discover the best-performing ML model for the task of cytotoxicity quantitative prediction, it is required to consider model accuracy, operation rate, and reasonable decision making, which is always a trade-off. Overall, 40 models were run on default parameters and tested on predictive accuracy (Table S1, Supporting Information), where top-5 performing models were Extra Trees, Histogram-based Gradient Boosting, light gradient-boosting machine library (LightGBM), Random Forest, and Bagging Regressors with negligibly different scores (**Figure 3**a). Since the top-5 models share similar performance, the choice is determined by time and decision-making process. The operation rate determines the robustness and applicability since the algorithm should rapidly process large amounts of data. This parameter should be considered and was measured for every model (Figure 3b). But it is decision making and model interpretability that are more important, therefore, we analyzed feature importance for three models, where Histogram-based

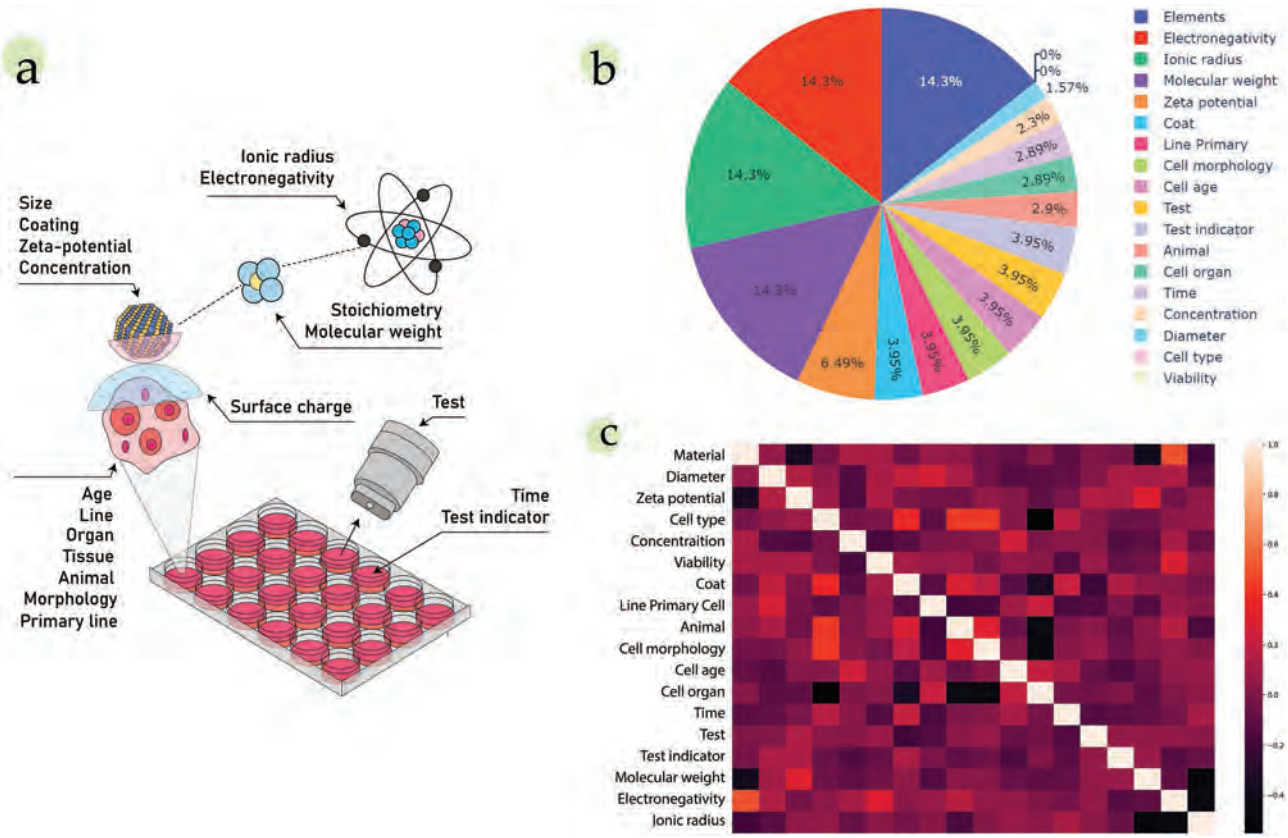


**Figure 1.** Data analysis: a) schematic view of model parameters, b) gaps statistics over the dataset variables, where the total number of single missing values is 51 666 out of 145 368 values in the initial dataset; c) correlation matrix on the processed dataset (two axes are identical, diagonal cells stand for feature self-correlation), 3087 samples.

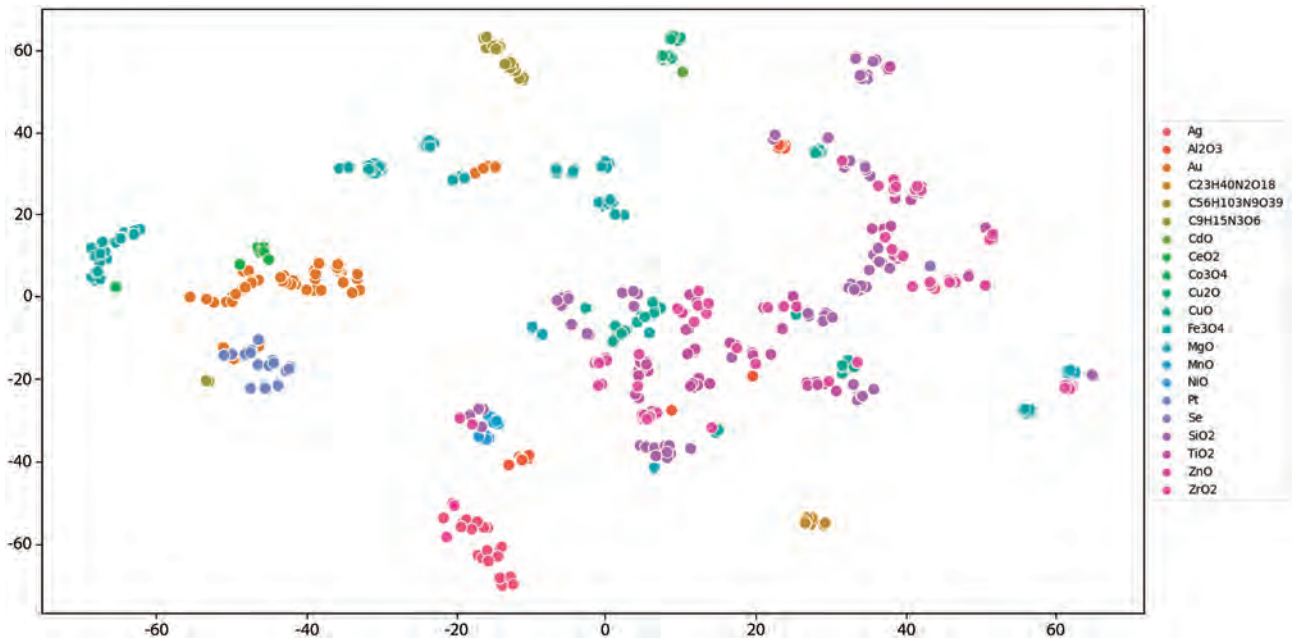


**Figure 2.** t-stochastic neighbor embeddings (t-SNE) clustering on full processed dataset (colors indicate different compounds).

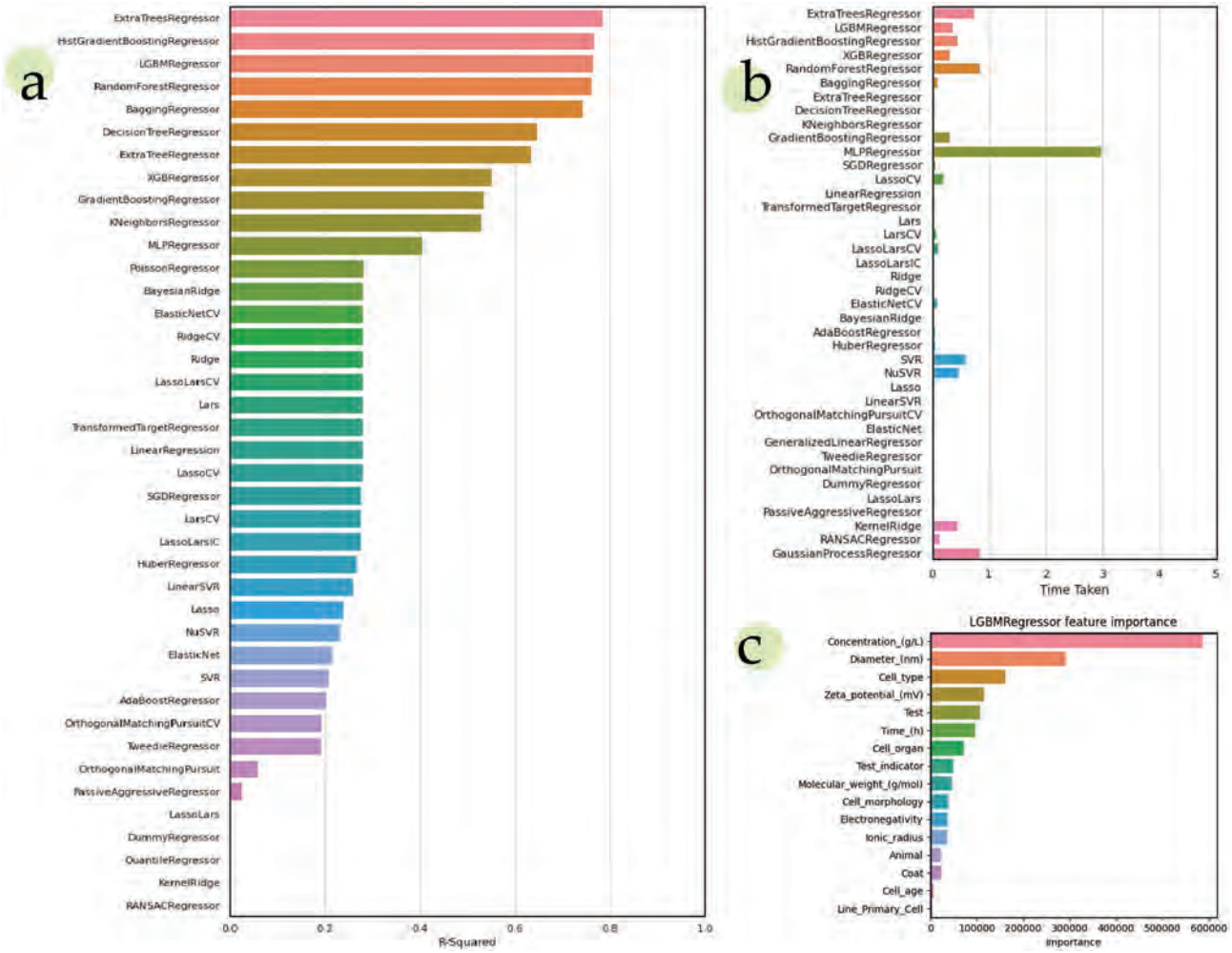


**Figure 3.** Model selection: a) 10-fold CV and b) time costs for the set of 40 screened ML models in minutes; c) feature importance of the final LGBM Regressor model.

Gradient Boosting and Bagging Regressors were discarded due to inability of feature importance determination. All these three models (Extra Trees, Random Forest, and LightGBM Regressors) produced intuitive results (Figure S6, Supporting Information) highlighting nanoparticle parameters such as concentration, diameter, zeta-potential, and compound molecular weight, and experiment conditions such as cell type, test type, and incubation time. As a result, the LightGBM model was chosen as the fastest (Figure 3c). Since all the models were run on default hyperparameters, the best model should be properly tuned. The LightGBM regressor was subjected to grid-search to tune all the hyperparameters, resulting in notable increase in 10-fold CV from 0.75 to 0.86 and RMSE from 16.1% to 12.2% (**Figure 4**a); hence, the final model is able to predict the cellular toxicity of inorganic nanoparticles with high accuracy.

What is more important for experimental scientists is the model's ability to produce experiment-related data, for instance, through understanding concentration cytotoxicity effects. Interestingly, the model was able to reproduce closely such effects for 100 nm negatively-charged $SiO_2$ nanoparticles in common MTT tests in the concentration range of 10–500 mkg $mL^{-1}$ correlating with the model uncertainty (Figure 4b). Even with these samples completely removed from the initial dataset (Figure 4c) the model preserved the reproducibility of these results showing its extrapolative power and successful knowledge learning. It can be concluded that the model did not only learn all the samples in the dataset but extracted meaningful patterns from it being able to transfer this knowledge to unseen samples.

Beyond the successful model itself, every AI approach should confirm current and generate new knowledge. Model SHAP values (Shapley additive explanations) were analyzed to reveal how the change of each numerical parameter influences the predicted viability value (Figure 4d). The SHAP values clearly show well-known concentration- and time-dependent effects, where increase in these parameters often leads to decreased cellular viability. As seen, nanoparticles with smaller hydrodynamic diameter more often led to increased viability, which is less intuitive due to the higher cellular internalization and surface area of smaller nanoparticles.[47] Nevertheless, much research demonstrated that the size effect cannot be easily isolated from the concentration effect, and bigger nanoparticles

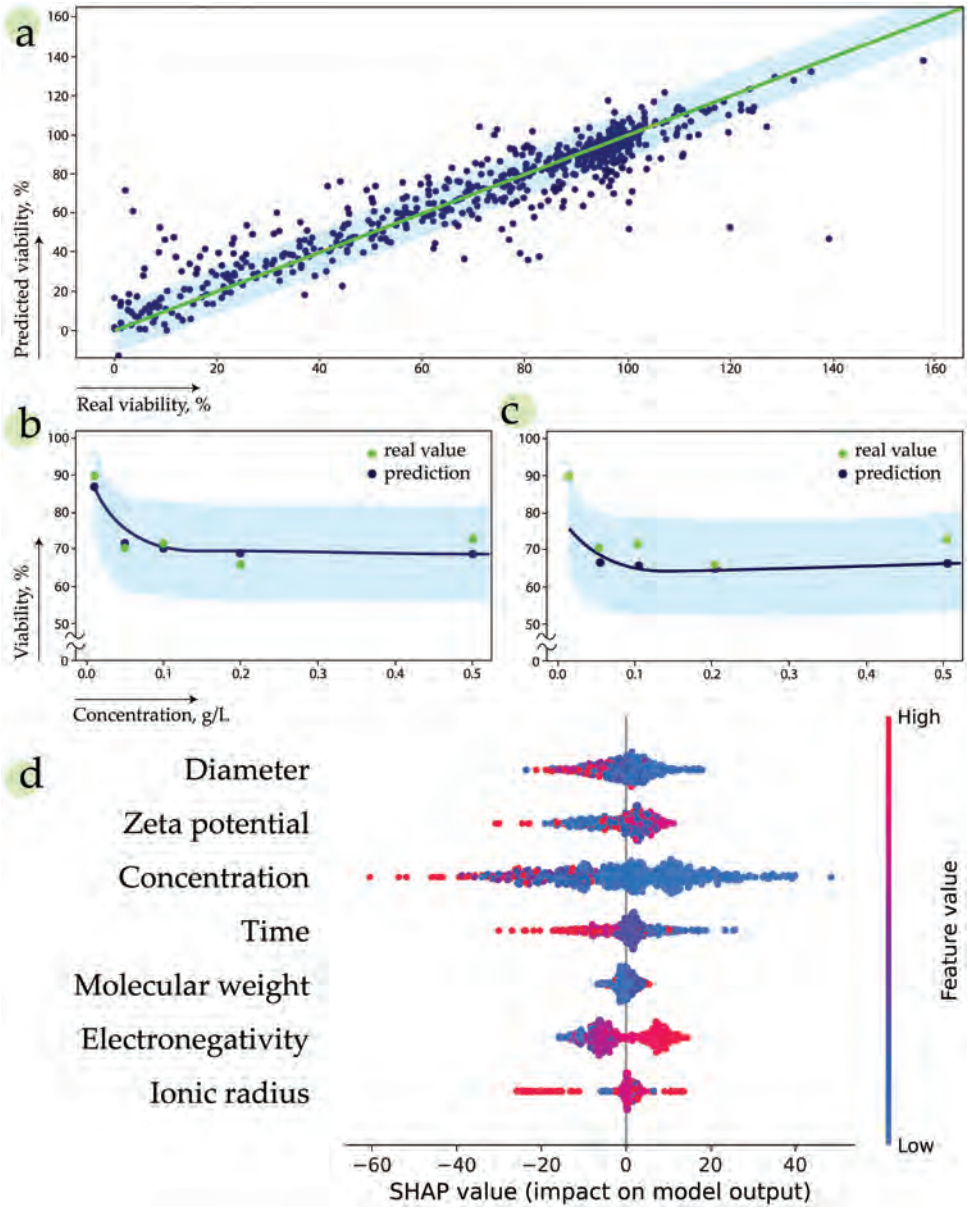


**Figure 4.** Final model performance a) on test data ($Q^2$ = 0.86 and RMSE = 12.2%), b) $SiO_2$ nanoparticles concentration plots reproduction on full dataset and c) on dataset with these samples excluded; d) SHAP model analysis on numerical dataset features. The light blue zone is limited by RMSE value cross-validation (CV)

often show greater toxicity in moderate and big concentrations. Interestingly, the SHAP analysis revealed strong element electronegativity influence on cellular viability, which can be explained by strong reductive properties and increased reactive oxygen species (ROS) production by materials containing elements with lower electronegativity leading to increased toxicity.[48] Therefore, the model produced knowledge consistent with the previous research and providing an intuition for non-cytotoxic nanoparticles design.

To evaluate the performance, advantages, and drawbacks of the developed approach, it was compared to the existing models closest to the topic (Table S4, Supporting Information). Binary classifiers do not allow to evaluate the extent of toxicity, whereas regression models predict nanoparticle concentrations showing predefined fixed viability (25% and 50% are observed in most cases) bearing no information about concentration dependencies. The developed model fulfills these gaps by working in wide concentration (from 0 to 1000 mkg $L^{-1}$) and viability (from 0% to 160%) ranges, moreover, predicting the latter directly. In addition, it spans over the widest range of cells (77 unique cell lines) and material types (21 unique compositions), allowing to study cytotoxicity for numerous applications. At the same time, the use of element-based nanomaterial descriptors gives the model ability to extrapolate on unseen compositions, where the current limitation is a chemical class, since the dataset consists almost entirely of metals and oxides. It should be noted that