令和 5 年度　博士学位論文

Applications of machine learning and literature data mining in drug discovery

(創薬における機械学習と文献データマイニングの応用)

大阪大学大学院　創成薬学専攻

医薬基盤科学講座・創薬情報科学

MARTIN

Table of Contents

Abbreviations

ADME, absorption, distribution, metabolism, and excretion;

AUC-ROC, area under the curve of the true-positive rate or recall;

CatBoost, categorical boosting;

CL_{int}, hepatic intrinsic clearance in liver microsome;

DNN, deep neural network;

DT, decision tree;

Extra Trees, extremely randomized trees;

F_a, fraction absorbed;

$f_{u,p}$, fraction unbound in plasma;

GB, gradient boosting;

ISO, International Standard Organization;

KNN, k-nearest neighbors;

LBDD, Ligand-based drug design;

LDA, linear discriminant analysis;

LightGBM, light gradient boosting machine;

LR, logistic regression;

nCV, Nested cross-validation;

RF, random forest;

PC1, Principal component 1;

SCCS, Scientific Committee on Consumer Safety:

SHAP, Shapley Additive exPlanations;

SMILES, simplified molecular-input line-entry system;

SiO_2-NPs, Amorphous silica nanoparticles;

SVM, support vector machine;

URAT1, urate transporter 1;

XGBoost, extreme gradient boosting;

3D, Three-dimensional;

3MRs, 3-membered rings;

Introduction

In drug discovery, conventional toxicological and biological methods, including both *in vivo* and *in vitro* approaches, are commonly used to evaluate the toxicity and activity of emerging nanoparticles or chemicals. However, relying solely on these techniques is considered less humane, efficient, and economical. Consequently, there is a growing demand for more timely risk assessment, cost-effective evaluation, and methods that minimize reliance on animal testing in both the pharmaceutical industry and health regulatory policies. In this context, *in silico* toxicity and activity predictions offer an alternative approach, providing cost-effective and efficient methods to rapidly assess whether specific nanoparticles or chemical compounds have the potential to pose adverse effects on human health or exhibit therapeutic utility. In particular, nanoparticles, those with a diameter of 100 nm or less, utilized in fields such as medicine and cosmetics, must undergo rigorous safety evaluations before they can be used for clinical translation. It has been noted that nanoparticles have the potential to penetrate cells and tissues through inhalation or skin contact. However, their complex behavior in biological environment has made it difficult to predict their potential toxicity.

This study explores the applications of machine learning and literature data mining in drug discovery. Machine learning, a subset of artificial intelligence, empowers computers to autonomously learn from data to make predictions. Literature data mining involves extracting implicit valuable information from a comprehensive dataset gathered from diverse individual studies, with a focus on addressing specific research questions where quantitative analysis of independently conducted experiments is predominant. Chapter 1 focuses on predicting cellular toxicity of amorphous silica nanoparticles, while Chapter 2 delves into predicting the activity of small molecules targeting urate transporter 1.

Main Paper

Chapter 1: Evidence-Based Prediction of Cellular Toxicity for Amorphous Silica Nanoparticles

Background

Decades of extensive research in nanotoxicology have yielded a wealth of data, prompting a shift towards literature data mining or meta-analysis to unveil hidden relationships within individual studies.[1–8] Unlike traditional approaches, which analyze nanoparticle toxicity using limited datasets based on specific attributes like particle size and concentration,[9–11] or omics-based biomarkers,[12] or predict other outcomes,[13,14] literature data mining integrates information from a global pool of evidence, enhancing generalizability across diverse experimental settings. This approach to developing data-driven models is particularly valuable for environmental and health-risk analyses.[1]

Previous literature data mining efforts have explored cellular toxicity for various nanoparticles such as cadmium-containing quantum dots,[1,2] carbon nanotubes,[3] graphene,[4] micro and nanoplastics,[5] nanoparticles,[6] phytosynthesized silvers,[7] and zinc oxides.[8] These studies formulated models based on experimental settings and physicochemical properties, employing cross-validation or split-sample internal validation. However, potential errors or biases in the collected literature data can compromise the reliability of these validations. Therefore, external validation, involving independently derived datasets, is crucial for ensuring the applicability of predictive models.[15–17]

When nanoparticles come into contact with proteins, such as those in serum, they form a layer of biomolecules called the corona on their surface. Thus, nanoparticle interactions with biological systems involve nanoparticle-corona complexes rather than pristine nanoparticles.[18] However, existing literature data mining reports[1–8] lack external validation or an assessment of the toxicological effects of preformed coronas in biological environments, limiting their real-world applicability. A cost-effective and rapid method is needed to develop a reliable prediction model for nanotoxicity.

Amorphous silica nanoparticles (SiO_2-NPs), extensively used in various industries,[19–22] including rubber, paints, cosmetics, biomedicine, and food additives, are regarded as a safety issue by the Scientific Committee on Consumer Safety (SCCS)[23] despite their large-scale production.[24–26] Hence, ensuring their safety is of utmost importance. *In vitro* cytotoxicity testing is an effective assessment of SiO_2-NP safety,[27,28] with smaller nanoparticles tend to induce greater toxicities.[29] Attributes like concentration, duration of exposure, surface chemistry, and synthetic pedigrees have the potential to influence the toxicity of SiO_2-NPs.[24–26] Despite numerous *in vitro* toxicological investigations, the key attributes contributing to the toxicity of SiO_2-NPs on a global scale still lack clarity.[24]

This study addresses the key attributes contributing to SiO_2-NP toxicity by proposing an evidence-based prediction method. Leveraging literature-mined SiO_2-NP cellular toxicity data, the method utilizes literature data mining, machine learning, and Shapley Additive exPlanations (SHAP) values[30] within a framework of nested cross-validation (nCV)[31] and internal and external validations. The resulting interpretable prediction model demonstrates satisfactory predictions and explanations for independent external toxicity data, proving the method's validity and reliability.

Methods

Figure 1.1 depicts the conceptual framework for an evidence-based approach to predict the toxicity of engineered nanoparticles, employing SiO_2-NPs as the experimental model and cytotoxicity as the measure of toxicity. The input attributes, encompassing SiO_2-NP physicochemical properties, experimental conditions, and cell types, along with binary output responses indicating cytotoxic or noncytotoxic outcomes, were initially gathered manually from the literature and organized in a tabular format. To establish a streamlined and cost-effective screening model for biocompatibility risk assessment, the cytotoxic responses were standardized using the International Standard Organization's (ISO) definition of cytotoxicity (ISO 10993-5), specifically classifying cytotoxicity (a positive label) as a reduction of more than 30% in cell viability.[27,28] All attributes were employed in the initial training of predictive model, utilizing various machine-learning algorithms to generate output responses based on input attributes. Subsequent utilization of SHAP values facilitated the identification of key attributes influencing SiO_2-NP toxicity. To validate the model's robustness and applicability, three essential validation steps were undertaken: nested cross-validation (nCV), internal validation, and external validation. In contrast to potentially optimistic estimates from non-nested cross-validation, nCV mitigates data leakage by incorporating an inner-loop CV within an outer CV. This dual CV structure is employed for both model selection, involving hyperparameter tuning through grid search in the inner loop, and model evaluation in the outer loop.[31]

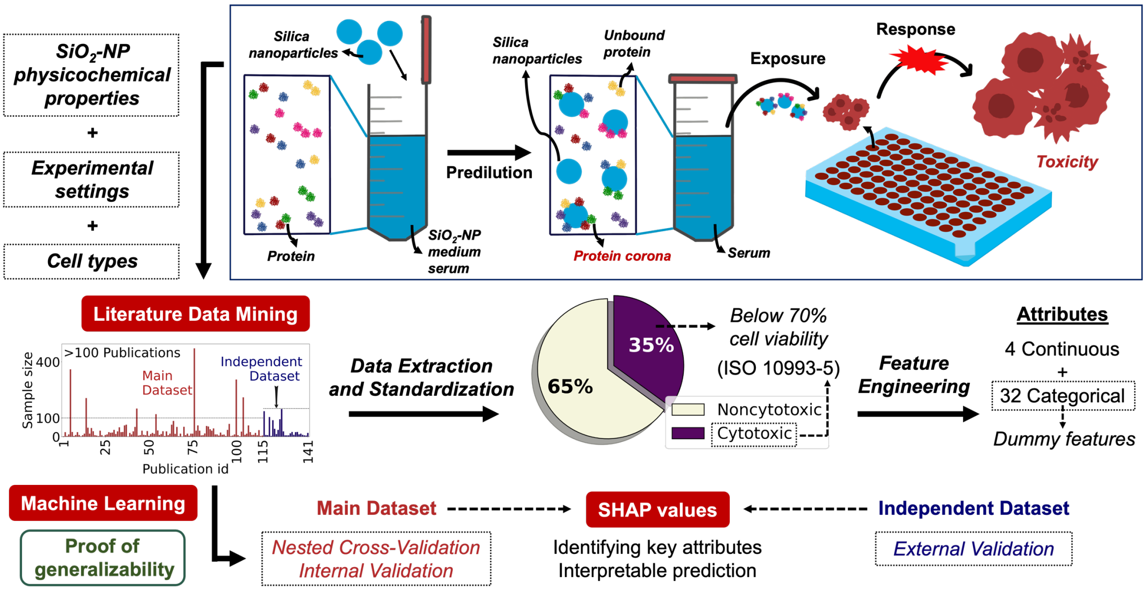


Figure 1.1. Framework of an evidence-based prediction method. *In vitro* cellular toxicity data were collected from published literature and standardized. Nested cross-validation, internal validation, and external validation were used to prove generalizability. Reprinted with permission from [*ACS Nano* 2023, 17, 11, 9987–9999]. Copyright [2023] American Chemical Society.

Literature data mining

Two comprehensive reviews on the toxicity of SiO_2-NPs served as the basis for the literature data evaluation in this study. These reviews, conducted by Napierska *et al.*[26] and Murugadoss *et al.*,[25] covered studies published up to 2010 and 2016, respectively. The selection of literature adhered to the PICOS framework[32] of evidence-based medicine, ensuring consistency and reliability across the chosen studies. The criteria for inclusion were: (1) a population involving human or mammalian cells; (2) the intervention and comparison focusing on amorphous non-mesoporous SiO_2-NPs vs. a negative control, with specified concentration, exposure time, and primary size ≤1000 nm; (3) the outcome centered on cytotoxicity (percentage of cell viability); and (4) the study design being an *in vitro* toxicological study. Exclusion criteria encompassed non-mammalian or co-cultured cells, crystalline or mesoporous SiO_2-NPs, abstract articles, and other non-relevant studies. Sixty-one studies meeting the inclusion criteria were identified from the two reviews, and their reference lists were reviewed for additional relevant literature, yielding 54 more eligible studies. In total, 115 studies were incorporated.

Systematic extraction of SiO_2-NP attributes and cell-viability data resulted in a main dataset comprising 4124 samples and 36 attributes. Mean cell viability values, obtained from text or graphs using WebPlotDigitizer,[33] were converted to binary labels: "1" (<70% cell viability, cytotoxic) and "0" (≥70% cell viability, noncytotoxic). The administered concentration of SiO_2 (in µg/mL for cell exposure) was utilized as the concentration attribute. Surface area (m^2/g) was calculated unless explicitly reported, with the formula: surface area = 6 ⁄ dr, where d is primary size in mm and r is density in g/cc. Due to missing data, categorical attributes with ranges were used for hydrodynamic size, zeta potential, and polydispersity index (PDI) attributes. Dummy features (binary vectors) were created for categorical attributes, with one dummy feature was omitted to avoid a dummy-variable trap (Supporting Information Tables S1). An attribute refers to a quantity that describes an instance in the dataset. Attributes have different types of domains, categorical (qualitative) and continuous (quantitative). Continuous attributes form a subset of real numbers (i.e., rational and irrational numbers), where there exists a discernible distinction among possible values. A feature is a specific value of an attribute, for example *Assay_viability* is an attribute; "*Assay_viability_MTT*" is a feature of the *Assay_viability* attribute. Feature scaling was applied using z-score normalization for linear and nonlinear-kernel classifiers and Min-Max normalization for deep neural network (DNN) classifier.

Machine learning

Thirteen established machine-learning algorithms were utilized (1) linear discriminant analysis (LDA), (2) logistic regression (LR), (3) ridge, (4) DNN, (5) k-nearest neighbors (KNN), (6) support vector machine (SVM), (7) decision tree (DT), (8) categorical boosting (CatBoost), (9) extremely randomized trees (extra trees), (10) gradient boosting (GB), (11) light gradient boosting machine

(LightGBM), (12) random forest (RF), and (13) extreme gradient boosting (XGBoost), categorized as linear (1-3), nonlinear (4), nonlinear kernels (5-6), and nonlinear tree-based classifiers (7-13). The scikit-learn (v1.0.2), tensorflow (v2.10.0), CatBoost (v1.0.4), LightGBM (v3.3.2), and XGBoost (v1.5.1) packages in Python 3.10 were employed for implementation.

The main dataset, containing 4124 samples, underwent shuffling and was split into training (80%) and internal test (20%) sets via random stratified sampling. Predictive models with binary classification were initially developed using 80% of the main dataset with all attributes, employing the 13 machine-learning algorithms. Ten-fold nCV was applied, and the models with a removed dummy feature were fine-tuned using GridSearchCV to obtain optimal hyperparameters (Supporting Information Tables S2–S3). Internal validation was conducted with the remaining 20% of the main dataset, independent of model building, and SHAP values were used to identify key attributes. Final predictive models were constructed using the entire main dataset and the identified key attributes

The evaluation metrics were based on accuracy ($1-\frac{TP+TN}{TP+FP+TN+FN}$), AUC-ROC (area under the curve of the true-positive rate or recall [$\frac{TP}{TP+FN}$] vs. false-positive rate [$\frac{FP}{FP+TN}$]), recall, and precision $\left(\frac{TP}{TP+FP}\right)$, where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively.

External validation

The final predictive models were employed to predict the independent datasets (905 samples) gathered from external studies published between 2017 and 2022[34–58] and in-house experiments. For instance, Gong *et al.* (2017)[34] investigated HaCaT cells exposed to 15-nm SiO_2-NPs (nine samples), while Liu *et al.* (2017)[35] studied A549 cells exposed to 15-nm SiO_2-NPs over 24 h (eight samples). Additionally, Nishijima *et al.* (2017)[36] examined the impact of 10–1000-nm SiO_2-NPs on THP-1 cells over 6–24 h (105 samples), and Premshekharan *et al.* (2017)[37] exposed THP-1 cells to 50-nm SiO_2-NPs for 22 h (four samples). Other studies (877 samples)[38–58] explored diverse SiO_2-NP sizes and exposure times on various cell lines, such as K17, HDF, LN229, N9, bEnd.3, HT-22, HEK293, hippocampal, HepG2, A549, SW480, HUVEC, GC-2spd, HeLa, BEAS-2B, Caco-2, H9c2, SH-SY5Y, NRK, BV2, L-02, and R28, among others.

Independent in-house experiments were conducted with 10-, 30-, 50-, 70-, 100-, 300-, and 1000-nm SiO_2-NPs (136 samples) sourced from Micromod Partikeltechnologie. The zeta potentials and hydrodynamic sizes of the SiO_2-NPs were measured using a Zetasizer Nano-ZS (Malvern Instruments Ltd.). The zeta potentials of the 10-, 50-, and 100-nm SiO_2-NPs were –15.6, –17.3, and –22.3 mV, respectively; their hydrodynamic sizes in water were 18.3, 48.4, and 99.8 nm, respectively. The zeta potentials and hydrodynamic sizes of 30-, 70-, 300-, and 1000-nm SiO_2-NPs were previously reported.[59,60] Exposure cell experiments were performed on A549, SH-SY5Y, TM4, BeWo, and RAW 264.7 cell lines, as detailed in Supporting Information Table S4. This comprehensive approach aimed to validate the robustness of the final predictive models across a spectrum of SiO_2-NP characteristics and experimental conditions.

Shapley Additive exPlanations (SHAP)

Attribute importance was established through global feature importance, as defined by the SHAP values:[30,61]

$\phi_i = \frac{1}{|F|!}\sum_{S\subseteq F\setminus\{i\}}|S|!\,(|F|-|S|-1)!\,[f_{S\cup\{i\}}(x_{S\cup\{i\}})-f_S(x_S)]$, where $F, S, x_S, f_{S\cup\{i\}}$, and f_S represent the set of all features, a subset of F, the values of the input features in the set S**,** a trained model with that feature present, and a trained model with that feature withheld, respectively. The SHAP value ϕ_i of the feature i was obtained by averaging the marginal contributions of all the permutations of a feature set. A greater mean absolute SHAP value signified a more influential feature in prediction. In this study, features displaying positive SHAP values directed the model output toward cytotoxicity, while those with negative values had the opposite effect, providing a rationale for decision-making in the predictive process.

Results

Literature data curation

We gathered cell viability data for 4124 samples, encompassing 32 categorical and 4 continuous attributes that characterize SiO_2-NP cellular toxicity. The data were sourced from 115 articles spanning the years 2004 to 2016, as depicted in Figure 1.2. Utilizing the ISO-10993-5 definition, 35% of the samples exhibited cytotoxic effects, while 65% were noncytotoxic. The attributes collected are detailed in Table 1.1, and their distribution is illustrated in Supporting Information Figure S1. Notably, our dataset, comprising 4124 samples and 36 attributes, surpassed the datasets of cadmium-containing quantum dots (3,028 samples, 24 attributes)[1,2] and nanoparticles (2,986 samples, 15 attributes)[6] by 36% and 38%, respectively. Moreover, our dataset exhibited a 50% and 140% greater diversity in attributes compared to the datasets of cadmium-containing quantum dots and nanoparticles, respectively.

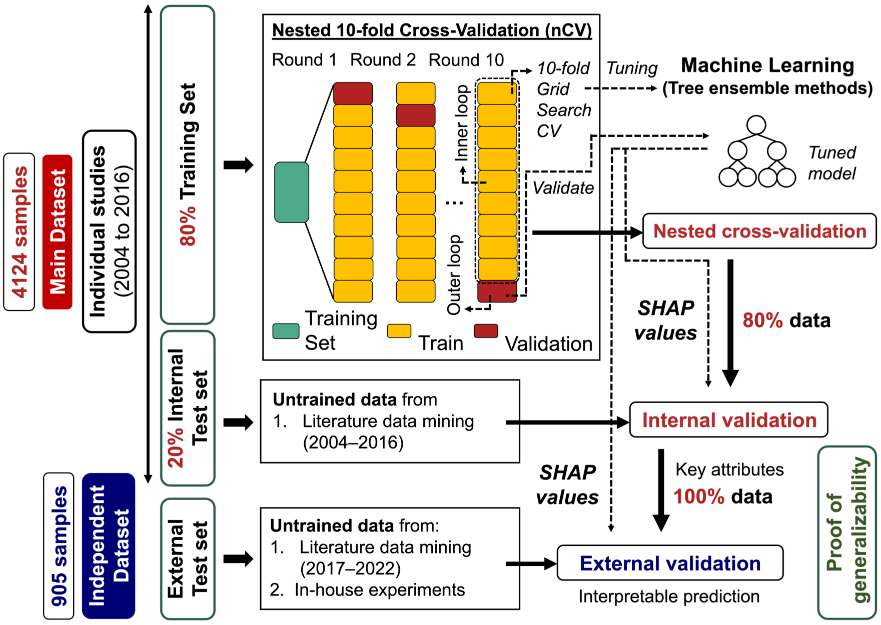


Figure 1.2. Data preparation: 80% of the main dataset containing all the attributes was trained and cross-validated using 10-fold nCV to develop the predictive model. The remaining 20% was used to internally validate the predictive model and identify the key attributes. Finally, 100% of the main dataset was used to build the final predictive model employing the identified key attributes to predict the independent dataset. Reprinted with permission from [*ACS Nano* 2023, 17, 11, 9987–9999]. Copyright [2023] American Chemical Society.

Table 1.1. Attributes of silica nanoparticles

No.	Attributes	Definition
SiO_2-NP Physicochemical Properties		
1	Primary_size	The average size of SiO_2 in the dry state measured by transmission electron microscopy (TEM), scanning electron microscope (SEM), or particle sizer.
2.	Primary_size_verification	The primary size of SiO_2 verified by the individual study, verified elsewhere (cited in previous publication), or not verified (directly used from manufacturer's specifications).
3.	Surface_area	The total area of SiO_2 surface measured by Brunauer–Emmett–Teller (BET) method or calculated by $^{6}/_{dr}$, where *d* is primary size in mm, *r* is density in g/cc.
4.	Hydrodynamic_size_water	The average hydrodynamic size of SiO_2 measured by dynamic light scattering in water.
5.	Hydrodynamic_size_culture	The average hydrodynamic size of SiO_2 measured by dynamic light scattering in culture medium.

Table 1.1. Continued

6.	Hydrodynamic_size_serum	The average hydrodynamic size of SiO_2 measured by dynamic light scattering in medium containing serum.
7.	Zeta_potential_water	The electrical potential of SiO_2 at the slipping plane or interface between SiO_2 surface and its water.
8.	Zeta_potential_PBS/HBSS	The electrical potential of SiO_2 at the slipping plane or interface between SiO_2 surface and its phosphate buffered saline (PBS) or Hank's balanced salt solution (HBSS).
9.	Zeta_potential_culture	The electrical potential of SiO_2 at the slipping plane or interface between SiO_2 surface and its culture medium.
10.	Zeta_potential_serum	The electrical potential of SiO_2 at the slipping plane or interface between SiO_2 surface and its medium containing serum.
11.	PDI_water	Polydispersity index (PDI), a measure of broadness of SiO_2 weight distribution in water.
12.	PDI_culture	Polydispersity index (PDI), a measure of broadness of SiO_2 weight distribution in culture medium.
13.	Surface_modification	The SiO_2 surface modifier, *e.g.*, chitosan, carboxyl, and amine.
14.	Surface_charge_water	The electrical charge of SiO_2 present at an interface in water.
15.	Surface_charge_culture	The electrical charge of SiO_2 present at an interface in culture medium.
16.	SiO_2-NP_synthesis	The SiO_2 synthetic pedigrees produced at high (*e.g.*, pyrolytic) or low (colloidal) temperature.
17.	SiO_2-NP_source	The source of SiO_2 obtained from in-house or commercial.
18.	SiO_2-NP_shape	The shape of SiO_2, either sphere or rod.
19.	SiO_2-NP_label	The label of SiO_2 including fluorescein-5-isothiocyanate (FITC), rhodamine, and iodine-125.
Experimental Settings		
20.	Concentration	A measured quantity of SiO_2 in μg/mL for exposure to cells.
21.	Exposure_time	The exposure duration of SiO_2 to cells.
22.	SiO_2-NP_medium_serum	The SiO_2 medium containing different serum concentrations (*e.g.*, serum-free, 10% fetal bovine serum [FBS], and bovine serum albumin [BSA]) for dilution or storage (prior exposure to cells).
23.	Assay_viability	An assay for measuring the cell viability, such as 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyl-2H-tetrazolium bromide or MTT.
24.	Viability_indicator	Cell viability indicator, *e.g.*, tetrazolium, lactate dehydrogenase (LDH), and adenosine triphosphate (ATP).

Table 1.1. Continued

25.	Viability_mechanism	Cell viability testing methods including structural cell damage, cell growth, and cellular metabolism.
26.	Interference_testing	The interference of SiO_2 with cell viability assay systems, either performed or not performed by the individual study.
27.	Positive_control	The use of positive control inducer, either included or not included by the individual study.
28.	Positive_control_inducer	A replicate containing all components of a test system and treated with a chemical/particle known to induce a positive response.
29.	Exposure_medium	The culture medium used during SiO_2 exposure to cells.
Cell Types		
30.	Cell_organ	Refers to organ or tissue from which cells originated.
31.	Cell_id	Identifies a specific cell, *e.g.*, A549, RAW 264.7, and HeLa.
32.	Cell_morphology	Refers to morphology of cells, mostly based on american type culture collection (ATCC), *e.g.*, epithelial, endothelial, and fibroblast.
33.	Cell_culture	The culture of cells, either primary cells (isolated from parental tissue) or cell lines (originated from primary cells).
34.	Cell_source	The source of cells including human, mouse, rat, pig, and hamster.
35.	Cell_age	The age of cells including embryonic and nonembryonic.
36.	Cell_disease	The disease stage of cells, either carcinoma or non-carcinoma.

Footnotes: Reprinted with permission from [*ACS Nano* 2023, 17, 11, 9987–9999]. Copyright [2023] American Chemical Society.

Nested cross-validation (nCV) and internal validation

We conducted nCV on 80% of the main dataset to obtain an initial unbiased assessment of predictive model accuracy, as depicted in Figure 1.2. Our analysis revealed that tree-based classifiers demonstrated a strong fit to the data, outperforming linear, DNN, and nonlinear kernel classifiers. Among these, CatBoost emerged as the top algorithm, achieving the highest nCV accuracy of 91.0±1.5%, as detailed in Table 1.2.

Primary evaluation metrics of the internal test set are accuracy and AUC-ROC (Table 1.2). Tree-based classifiers demonstrated satisfactory accuracies ranging from 85.6% to 90.4% and excellent AUC-ROCs between 94.1% and 96.3% (except for DT, with 86.0%). Linear, DNN, and nonlinear kernel classifiers exhibited accuracies of 75.2% to 84.6% and AUC-ROCs of 82% to 89.8%. CatBoost consistently outperformed other algorithms, achieving an accuracy of 90.4%, an AUC-ROC of 96.3%, recall of 85.6%, and precision of 87.1%.

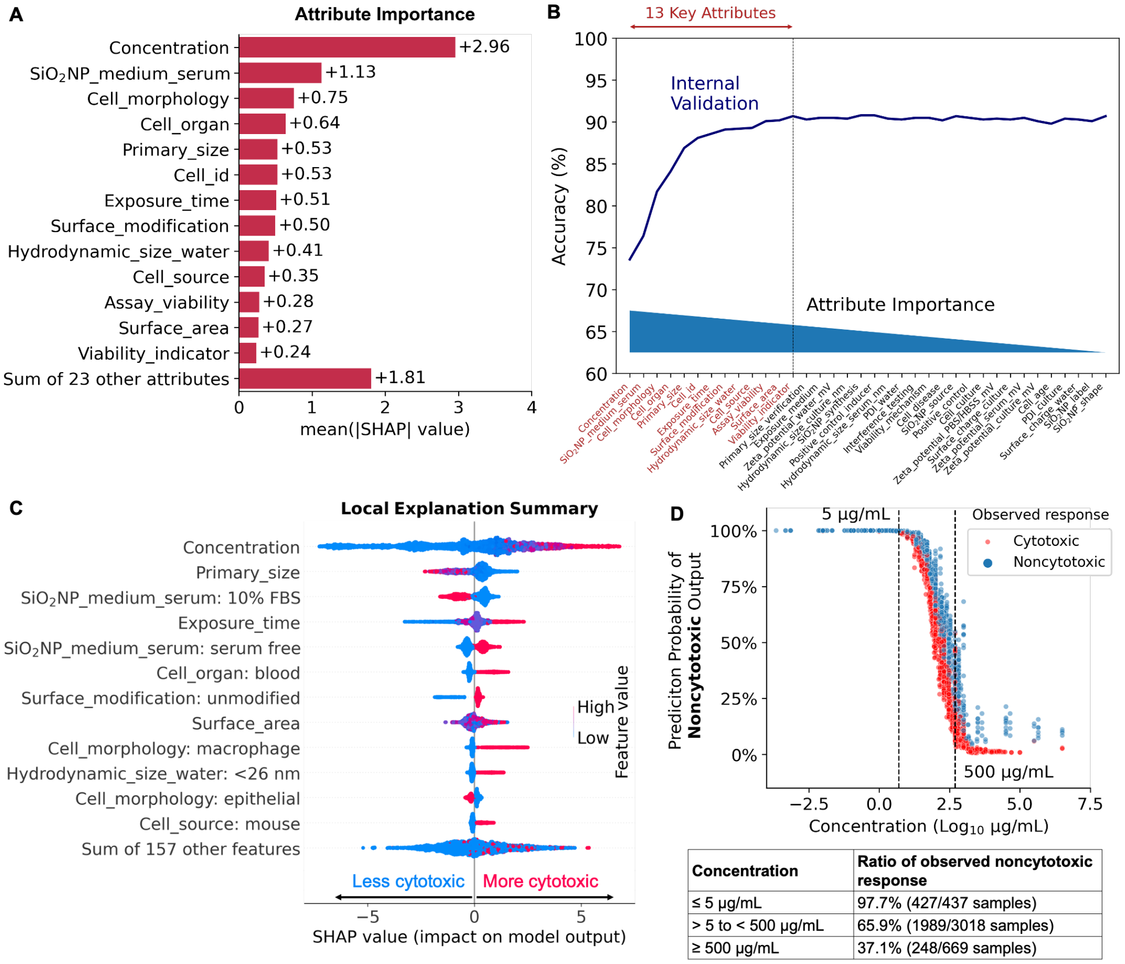
Table 1.2. Prediction-error comparisons: Internal validation (All 36 attributes and 824 samples)

Machine Learning	$nCV_{10\text{-fold}}$	Accuracy	AUC-ROC	Recall	Precision
Linear					
LDA	74.5±2.5%	75.2%	82.6%	56.2%	68.0%
LR	82.3±1.9%	83.4%	89.8%	73.3%	78.4%
Ridge	75.4±2.3%	75.3%	82%	52.7%	70.0%
Nonlinear					
DNN	75.2±1.7%	75.7%	82.7%	61.8%	67.1%
Kernel					
KNN	85.3±1.8%	84.6%	82.5%	75.3%	80.0%
SVM	84.3±1.8%	83.0%	87%	70.5%	79.2%
Tree-based					
DT	87.3±1.8%	85.6%	86.0%	81.5%	78.5%
Extra Trees	86.9±1.8%	85.6%	94.1%	76.7%	81.5%
RF	88.1±1.9%	87.4%	94.5%	79.5%	84.1%
CatBoost	91.0±1.5%	90.4%	96.3%	85.6%	87.1%
GB	90.3±2.0%	89.1%	95.3%	83.6%	85.3%
LightGBM	90.0±1.6%	90.1%	95.8%	84.9%	86.7%
XGBoost	90.2±1.7%	89.9%	95.8%	84.9%	86.4%

Footnotes: LDA, linear discriminant analysis; LR, logistic regression; DNN, deep neural network;

KNN, k-nearest neighbors; SVM, support vector machine; DT, decision tree; Extra Trees, extremely randomized trees; RF, random forest; CatBoost, categorical boosting; GB, gradient boosting; LightGBM, light gradient boosting machine; XGBoost, extreme gradient boosting. Reprinted with permission from [*ACS Nano* 2023, 17, 11, 9987–9999]. Copyright [2023] American Chemical Society.

We utilized attribute importance for feature selection via SHAP values with CatBoost. Based on the attribute importance (Figure 1.3A), we identified the top 13 attributes that resulted to optimal predictive accuracy (Figure 1.3B), arranging them in order of importance: *concentration, SiO_2-NP_medium_serum, cell_morphology, cell_organ, primary_size, cell_id, exposure_time, surface_modification, hydrodynamic_size_water, cell_source, assay_viability, surface_area,* and *viability_indicator* (refer to Table 1.1 and Supporting Information Figure S1).



Concentration	Ratio of observed noncytotoxic response
≤ 5 μg/mL	97.7% (427/437 samples)
> 5 to < 500 μg/mL	65.9% (1989/3018 samples)
≥ 500 μg/mL	37.1% (248/669 samples)

Figure 1.3. Attribute importance for silica nanoparticles, based on CatBoost. (A) Global interpretability for the average absolute SHAP value magnitudes. (B) Predictive accuracy of internal validation with incrementally increasing attributes. (C) Local interpretability, with each dot corresponding to a sample of silica nanoparticle cellular toxicity obtained from 100% of the main dataset. (D) The prediction probability of CatBoost to output a noncytotoxic class at a given condition of concentration attribute alone, using 100% of the main dataset. Reprinted with permission from [*ACS Nano* 2023, 17, 11, 9987–9999]. Copyright [2023] American Chemical Society.

Subsequently, we reconstructed the predictive models using 80% of the main dataset and the identified key attributes and assessed their performance using the internal test set (Table 1.3). Instead of employing all attributes (as shown in Table 1.2), comparable performance was achieved using solely the 13 key attributes, with CatBoost exhibiting the best performance (accuracy: 90.7%, AUC-ROC: 95.9%, recall: 85.6%, precision: 87.7%, and nCV: 90.3±1.9%). Other tree-based classifiers, including RF, GB, LightGBM, and XGBoost, also demonstrated high scores (accuracy >88%, AUC-ROC >94%, recall >81%, precision >85%, and nCV >88%).

Table 1.3. Prediction-error comparisons: Internal validation (13 key attributes and 824 samples)

Machine Learning	nCV$_{10\text{-fold}}$	Accuracy	AUC-ROC	Recall	Precision
Linear					
LDA	74.1±2.2%	73.9%	80.4%	48.6%	68.6%
LR	74.7±2.1%	73.3%	80.2%	45.5%	68.6%
Ridge	74.4±2.1%	73.9%	80%	46.9%	69.5%
Nonlinear					
DNN	74.2±2.9%	76.3%	83.7%	67.3%	66.4%
Kernel					
KNN	85.1±1.9%	85.2%	82.8%	74.7%	82.0%
SVM	85.2±1.9%	85.2%	89%	73.3%	82.9%
Tree-based					
DT	86.3±1.5%	87.3%	86.2%	81.2%	82.6%
Extra Trees	86.5±2.0%	86.1%	94.1%	77.1%	82.4%
RF	88.1±2.0%	88.8%	94.9%	81.2%	86.5%
CatBoost	90.3±1.9%	90.7%	95.9%	85.6%	87.7%
GB	89.4±2.0%	89.0%	95.1%	83.2%	85.3%
LightGBM	88.5±1.6%	89.1%	95.1%	82.9%	85.8%
XGBoost	89.4±1.5%	89.7%	95.4%	83.9%	86.6%

Footnotes: LDA, linear discriminant analysis; LR, logistic regression; DNN, deep neural network; KNN, k-nearest neighbors; SVM, support vector machine; DT, decision tree; Extra Trees, extremely randomized trees; RF, random forest; CatBoost, categorical boosting; GB, gradient boosting; LightGBM, light gradient boosting machine; XGBoost, extreme gradient boosting. Reprinted with permission from [*ACS Nano* 2023, 17, 11, 9987–9999]. Copyright [2023] American Chemical Society.

Finally, we constructed the final predictive models using the 13 key attributes from 100% of the main dataset (4124 samples) and obtained a robust nCV accuracy (Table 1.4). Next, we examined how the SHAP values were distributed for the 13 key attributes across the various samples. In accordance with the SHAP local explanation summary (Figure 1.3C), a larger SiO_2-NP primary size, the presence of 10% fetal bovine serum (FBS) in the SiO_2-NP medium (prior exposure to cells), surface-modified SiO_2-NPs, and cells with epithelial morphologies were associated with reduced cytotoxic effects. In contrast, a higher concentration of SiO_2-NPs, an extended exposure time and surface area, a hydrodynamic size less than 26 nm in water, the absence of serum in the SiO_2-NP medium, and the presence of mouse cells, macrophage cells, blood cells, and a tetrazolium viability indicator with an MTT assay (Supporting Information Figure S2) were linked to increased cytotoxicity. MTT assay is highlighted as it exhibits the highest ranking in terms of the *Assay_viability* attribute, as indicated in Supporting Information Figure S2, offering a detailed overview of the local explanation not replicated by other viability assays. While concentration emerged as a leading attribute determining SiO_2-NP toxicity, SiO_2-NPs with concentrations >5 μg/mL alone did not ensure accurate prediction, as depicted in Figure 1.3D. Remarkably, 97.7% of SiO2-NPs with concentrations ≤5 μg/mL were linked to noncytotoxicity. Clear thresholds were not observed for other continuous attributes (see Supporting Information Figure S3). Furthermore, a singular decision tree with an nCV accuracy of 73.4±1.9%

was identified (refer to Supporting Information Figure S4) for simplified guidance on SiO_2-NP toxicity. Nevertheless, for optimal predictive efficacy, we advise utilizing all 13 key attributes when employing our model *via* Google Colab (https://github.com/martinj-phs/nanosilica).

External validation

We created an independent dataset comprising 905 samples, distinct from the main dataset, adding complexity, value, and real-world relevance to the task of predicting and explaining SiO_2-NP toxicity. External validation results (Table 1.4) revealed that CatBoost exhibited satisfactory generality and yielded the highest performance (accuracy: 88.1%, AUC-ROC: 92.0%, recall: 72.4%, and precision: 78.0%), followed by GB, RF, and XGBoost (accuracies >84% and AUC-ROCs >88%). Notably, RF displayed the lowest recall (48.4%) among tree-ensemble classifiers, making it unsuitable for identifying all positive samples, unlike boosting algorithms (CatBoost, GB, XGBoost, and LightGBM) with recall rates exceeding 61%. Linear, DNN, nonlinear kernel, and DT classifiers struggled to fit the independent dataset, achieving accuracies between 64.4% and 75.9%. SVM solely predicted the majority noncytotoxic class, exhibiting the poorest AUC-ROC (46%) and recall (3.1%), indicating frequent misclassification and failure to identify positive samples.

To assess the impact of serum in predicting SiO_2-NP toxicity, we reconstructed predictive models using 12 key attributes, excluding the *SiO_2-NP_medium_serum* attribute. Overall, the results indicated significantly reduced performance (CatBoost: accuracy, 80.7%; AUC-ROC, 84.4%; recall, 53.3%; precision, 63.2%; and nCV, 88.7±1.3%), emphasizing the critical role of nanoparticle-corona formation in biologically diverse environments containing varying serum concentrations for highly accurate predictions (Supporting Information Table S5). Additionally, Supporting Information Figure S5 demonstrated lower model performance when all 36 attributes were used, underscoring the importance of attribute selection to prevent overfitting in a truly independent test set.

Table 1.4. Prediction-error comparisons: External validation (13 key attributes and 905 samples)

Machine Learning	$nCV_{10\text{-fold}}$	Accuracy	AUC-ROC	Recall	Precision
Linear					
LDA	73.6±2.4%	65.2%	70.2%	64.4%	38.2%
LR	74.4±2.5%	64.4%	64.1%	57.8%	36.4%
Ridge	74.3±1.8%	65.3%	70%	63.6%	38.1%
Nonlinear					
DNN	75.3±2.1%	65.5%	68.1%	52.4%	36.3%
Kernel					
KNN	86.5±1.4%	74.0%	71.7%	67.1%	48.4%
SVM	86.3±2.1%	75.9%	46%	3.1%	100.0%
Tree-based					
DT	87.7±1.6%	67.4%	59.7%	44.0%	36.9%
Extra Trees	87.5±1.8%	82.3%	88.4%	57.8%	66.7%
RF	88.7±1.6%	85.1%	91.4%	48.4%	85.2%
CatBoost	90.5±1.6%	88.1%	92.0%	72.4%	78.0%
GB	89.8±1.4%	87.8%	90.2%	66.2%	81.4%
LightGBM	89.3±1.3%	82.0%	88.1%	67.6%	62.8%
XGBoost	89.6±1.4%	84.5%	88.4%	61.3%	72.3%

Footnotes: LDA, linear discriminant analysis; LR, logistic regression; DNN, deep neural network;

KNN, k-nearest neighbors; SVM, support vector machine; DT, decision tree; Extra Trees, extremely randomized trees; RF, random forest; CatBoost, categorical boosting; GB, gradient boosting; LightGBM, light gradient boosting machine; XGBoost, extreme gradient boosting. Reprinted with permission from [*ACS Nano* 2023, 17, 11, 9987–9999]. Copyright [2023] American Chemical Society.

Complex Relationships of SiO_2-NP Attributes with Cellular Toxicity

CatBoost was chosen to represent the prediction outcomes for external validation. We employed hierarchical clustering to group the independent datasets based on their similarity in explanation (SHAP values), visualizing heterogeneity (Figure 1.4A). Prediction errors for 905 samples (55 sets of experiments) are presented in Supporting Information Figures S6-S7, with two representative sets detailed in Figures 4B and 4C. Decision plots for correctly classified and misclassified samples can be found in Supporting Information Figure S8. To ensure real-world applicability, we employed SHAP values to quantitatively elucidate the CatBoost process generating the output cellular toxicity response from input key attributes. Figures 1.4D–G and Supporting Information Rationality depict the rational decision-making and complex attribute relationships governing potential SiO_2-NP hazards and their impact on cellular machinery.

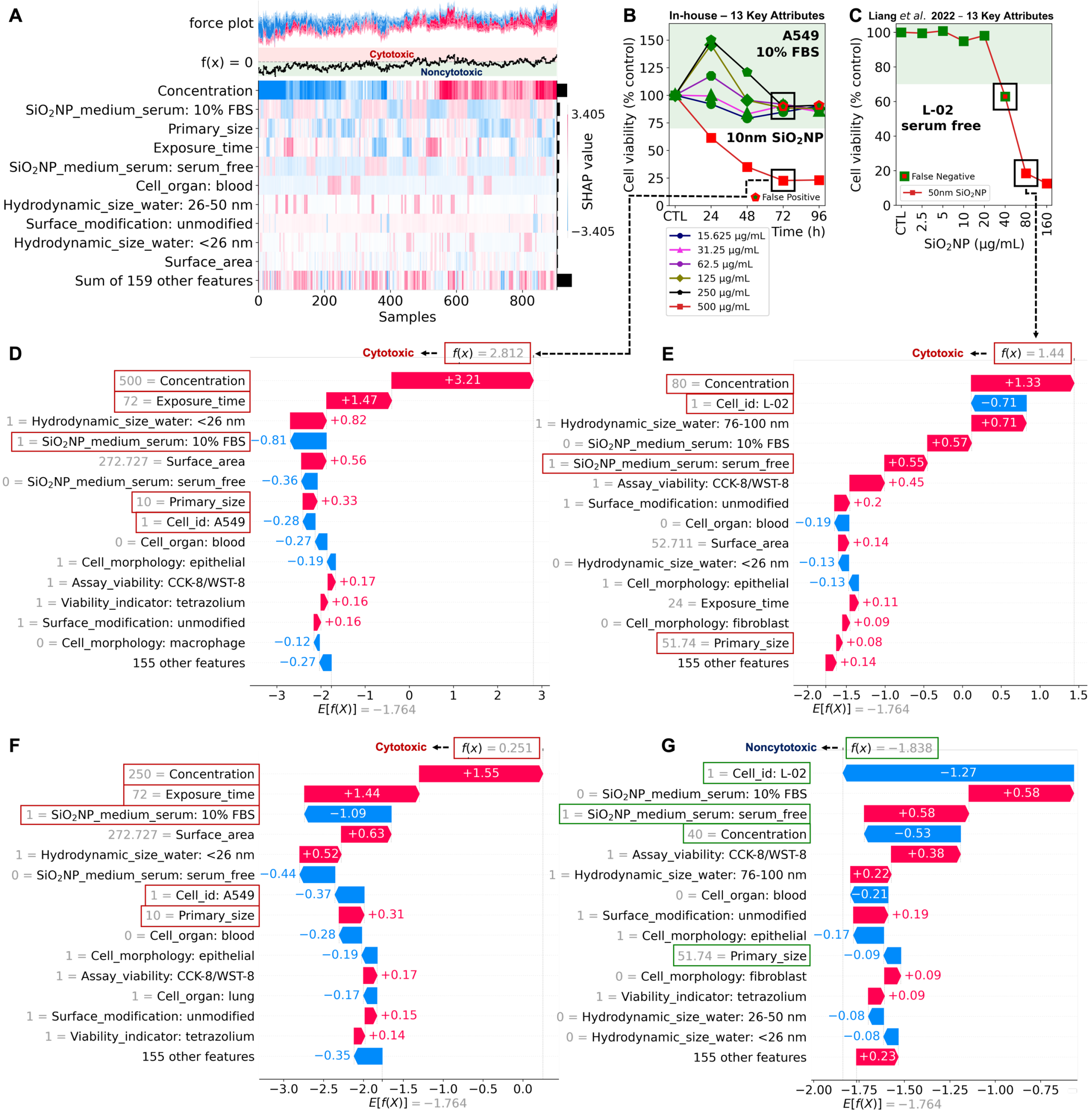


Figure 1.4. Prediction errors generated by the CatBoost model upon external validation. (A) SHAP heatmap plot. Samples with similar SHAP-value-based explanations were grouped together *via* hierarchical clustering. Increasing and decreasing cytotoxicity by attribute value are indicated in red and blue, respectively. The force plot at the top corresponds to the ratios of attribute values with a negative magnitude (blue) to those with a positive magnitude (red); f(x) = 0 corresponds to the predicted cytotoxicity. Samples predicted to be cytotoxic and noncytotoxic are shown in the red and green regions, respectively. (B and C) Prediction errors of each sample from two of the 55 sets of experiments. Red and green markers indicate cytotoxicity and noncytotoxicity, respectively. Correctly

classified samples have either a green or red marker, whereas misclassified samples have markers that are a combination of both colors. (D and E) Two examples of correctly classified samples. The positive values of f(x) = 2.812 and f(x) = 1.44 correspond to the cytotoxic class and were generated from the sum of the base value (–1.764) and the additive contributions of each attribute value (3.21 + 1.47 + …. – 0.27 in f(x) = 2.812 and 1.33 – 0.71 + …. + 0.14 in f(x) = 1.44). They explain which attribute value corresponded to the predicted cytotoxicity values of 2.812 and 1.44 from the base value; for example, in f(x) = 2.812, *concentration: 500 µg/mL* increased the base value by 3.21, whereas *SiO_2-NP_medium_serum: 10%_FBS* decreased it by 0.81. The base value was the average cytotoxicity value of the entire main dataset. (F and G) Two examples of misclassified samples. The positive and negative values of f(x) = 0.251 and f(x) = –1.838 correspond to the cytotoxic and noncytotoxic class, respectively. Reprinted with permission from [*ACS Nano* 2023, 17, 11, 9987–9999]. Copyright [2023] American Chemical Society.

Discussion

Differentiating between cytotoxic and noncytotoxic nanoparticles is crucial for nanosafety. The CatBoost model, derived from a comprehensive literature data mining effort covering 115 publications, unveiled key SiO_2-NP attributes essential for predicting toxicity. These attributes, including *concentration, SiO_2-NP_medium_serum, cell_morphology, cell_organ, primary_size, cell_id, exposure_time, surface_modification, hydrodynamic_size_water, cell_source, assay_viability, surface_area,* and *viability_indicator*, formed the basis of an evidence-based prediction model for SiO_2-NP toxicity. The dataset, comprising 36 diverse attributes and 4124 samples, along with an independent dataset of 905 samples, constituted the largest and most comprehensive set to date.

While previous literature data mining efforts[1–8] failed to recognize the rapid formation of protein coronas around nanoparticles in biological fluids,[62] our study emphasized the importance of considering the biological medium attribute, particularly the *SiO_2-NP_medium_serum* attribute, for accurate nanotoxicity predictions. The absence of this attribute led to a significant drop in predictive performance. Our findings challenge existing nanotoxicity models and underscore the necessity of accounting for preformed coronas in biological environments for successful predictive modeling.

Cellular uptake of SiO_2-NPs triggers underlying mechanisms related to concentration, time, size, surface, cell, and serum attributes.[43,63–67] A higher concentration of SiO_2-NPs results in a proportional increase in the amount adsorbed by cells and higher internalization efficiency.[65] SiO_2-NP concentrations below 5 μg/mL are linked to noncytotoxicity, possibly due to the negligible uptake by cells at these low levels.[65] At these levels, SiO_2-NP may not interact extensively with the cellular components, and their impact on cell viability is minimal, highlighting the importance of concentration levels in determining the potential harm of SiO_2-NPs. Extended exposure time enhances the efficiency of SiO_2-NP internalization into cells.[64] The surface area plays a crucial role, as increasing the size of SiO_2-NPs up to 50 nm reduces the total surface area, thereby preventing internalization. Moreover, SiO_2-NPs with a hydrodynamic size less than 26 nm in water demonstrate better internalization efficiency.[43] The absence of serum in the SiO_2-NP medium strengthens the adhesion of SiO_2-NPs to cell membranes, increasing internalization.[63–66] The presence of various cells, particularly nonphagocytic cells, exhibits lower efficiency in endocytosis compared to phagocytic monocytes/macrophages, possibly due to their larger size, leading to exclusion from developing pinocytic vesicles.[67] The choice of a viability indicator introduces the possibility of SiO_2-NPs interacting with the assay, contributing to the overall understanding of SiO_2-NP toxicity.

The role of the serum attribute in predicting SiO_2-NP toxicity is evident; preformed coronas in the presence of serum has the potential to alleviate SiO_2-NP toxicity. Corona formation alters cell receptor recognition of SiO_2-NPs and, by reducing SiO_2-NP surface energy, hinders efficient interaction of surface silanols [≡Si–OH and =Si(OH)$_2$] with biomembranes, thereby lowering SiO_2-NP uptake efficiency.[64–66] However, the absence of serum can lead to more cytotoxic effects, as surface silanols of SiO_2-NPs can directly engage with and disturb cellular membranes through hydrogen bonding and electrostatic interactions. The scientific reason behind the reduction in surface energy of nanoparticles when surrounded by a biomolecular corona lies in the interactions between the nanoparticles and biomolecules. The biomolecular corona formed around nanoparticles reduces surface energy by providing steric stabilization, shielding from direct exposure, facilitating biological recognition, and passivating the surface.[18] Notably, a specific surface-silanol pattern known as "nearly free silanol" facilitates membranolysis by interacting with phosphatidylcholine, supporting the idea that surface modification can reduce SiO_2-NP toxicity, irrespective of silica crystallinity.[68]

Evidence indicates that nanoparticles produced at high temperatures (pyrolytic) might be more toxic.[24] Nanoparticles treated at high temperatures exhibit strained 3-membered rings (3MRs) on their surface. The strained nature of 3MRs makes them prone to homolytic cleavage, generating hydroxyl radicals upon water adsorption. This structural feature enhances hydrolysis compared to unstrained siloxane bonds, resulting in the formation of nonhydrogen-bonded hydroxyl groups when exposed to water vapor.[69] However, the *SiO_2-NP_synthesis* attribute did not emerge as a key attribute. This may be attributed to the fact that only a single study directly compared pyrolytic and colloidal SiO_2-NPs with varying synthetic pedigrees under identical conditions.[70] This underscores the necessity for more in-depth investigations into the impact of synthetic pedigrees on SiO_2-NP toxicity, taking into account variations in size, surface, cell, assay, and biological media.

External validation is essential for implementing highly accurate generalizations in real-world scenarios.[15–17] The CatBoost model consistently exhibited satisfactory performance for both internal validation (accuracy: 90.7%, AUC-ROC: 95.9%, recall: 85.6%, and precision: 87.7%; nCV: 90.3±1.9%) and external validation (accuracy: 88.1%, AUC-ROC: 92.0%, recall: 72.4%, and precision: 78.0%; nCV: 90.5±1.6%). Thus, CatBoost emerged as a more promising algorithm for nanotoxicity generalizability compared to the previously used RF or DT.[1,3–8] The unexpected poor performance of DT and kernel classifiers in external validation despite favorable internal validation results underscores the pivotal role of thorough external validation.

CatBoost's outperformance over RF can be attributed to its sequential learning strategy, where multiple trees are trained one stage at a time, correcting errors from previous fits. This sequential approach allows CatBoost to effectively capture complex patterns in the data, enhancing its ability to correct errors and improving overall predictive accuracy. In contrast, RF constructs trees independently, potentially missing intricate relationships in the data. The diversity among trees in RF may yield conservative decision boundaries, prioritizing overall accuracy, but potentially resulting in lower recall. CatBoost's built-in support for categorical features eliminates the need for manual encoding, offering a notable advantage in nanoparticle datasets with categorical variables. Additionally, CatBoost incorporates regularization techniques to prevent overfitting, enabling robust generalization to unseen data. CatBoost also handles missing data naturally, a valuable feature in real-world datasets where missing values are common. In summary, the sequential learning process, categorical features handling, regularization, and handling in missing data in CatBoost provide advantages over RF.

Despite the comprehensive nature of our study, some limitations should be acknowledged. The current predictive model does not provide quantitative values for the extent of SiO_2-NP surface energy reduction associated with changes in concentration, serum, exposure time, and size, nor does it quantify the resulting impact on toxicity. This aspect presents a potential avenue for future research, and efforts can be directed toward developing models that quantitatively assess the relationship between these variables and the reduction in surface energy, along with its implications for toxicity outcomes. The choice of attributes, such as using administered concentration instead of cellular dose or number of particles, reflects data availability constraints. Future research may benefit from incorporating these parameters if obtainable. Additionally, a more exhaustive characterization of SiO_2-NP physicochemical properties is warranted. Furthermore, the development of tools capable of automatically extracting nanoparticle data in a high-throughput manner would be extremely advantageous in the future.

In vitro findings presented in this study may not directly extrapolate to *in vivo* outcomes, emphasizing the challenge of establishing *in vitro-in vivo* correlation. To enhance the predictive power of models, future studies should consider using specific types of nanoparticles and avoid making exaggerated claims about nanotoxicity predictions without adequate external validation.

The evidence-based method offers a promising framework for nanotoxicological research, incorporating global evidence to develop reliable predictive models. A frequently employed ratio to evaluate the models in real world-practice is 80:20, signifying that 80% of the data is allocated for training (4124-sample of main dataset), while the remaining 20% (905-sample of independent dataset) is designated for testing. The SiO_2-NP case study illustrates the applicability of the method, providing insights into key attributes influencing SiO_2-NP toxicity. The CatBoost[71] model, employed as an effective tool for nanotoxicity prediction, demonstrates quantitative interpretability in generating cytotoxicity responses from key attributes. External validation proves crucial for ensuring the model's generalizability. We anticipate that our integrated approach, uniting literature data mining, machine learning, and SHAP values, can serve as a versatile platform in examining various engineered nanoparticles for predicting and explaining diverse biological outcomes. This study has the potential to advance the development of safe nanoparticles for biomaterials and provides reliable guidance for predictions in nanoinformatics.

Supporting Information

Final model code, Nanosilica dataset, Supporting information (Tables S1-S5; Fig. S1-S8), and Supporting information rationality to this chapter 1 can be found online at http://dx.doi.org/ 10.1021/acsnano.2c11968. Reprinted with permission from [*ACS Nano* 2023, 17, 11, 9987–9999]. Copyright [2023] American Chemical Society.

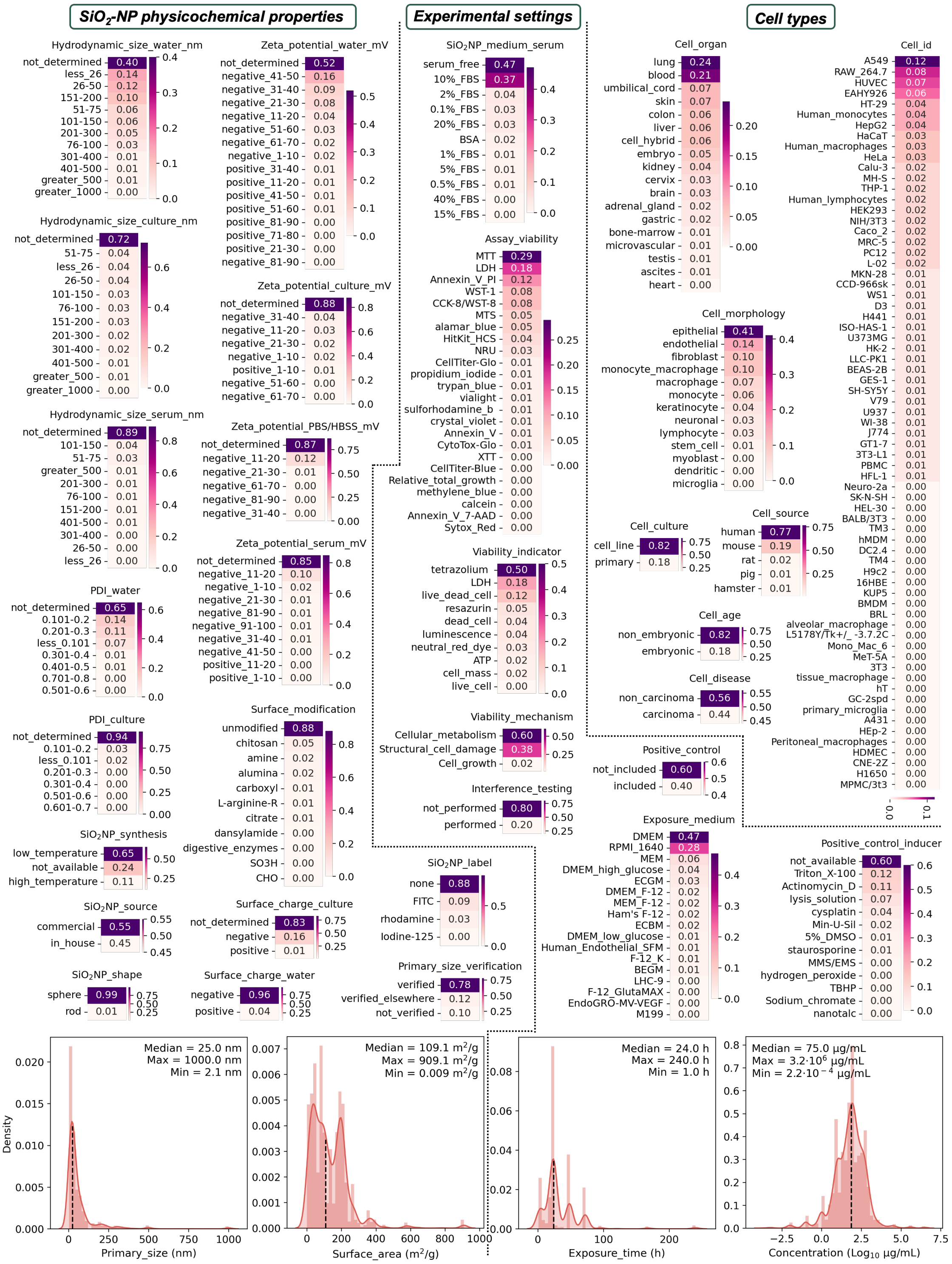


Figure S1. Silica nanoparticles with 32 categorical (heatmap visualization) and 4 continuous (distribution plot visualization) attributes. The distribution of attributes can be read as follows, for

example in SiO_2-NP_synthesis attribute, silica nanoparticles were synthesized at low temperatures (e.g., sol-gel) for 65% of 4124 samples and at high temperatures (e.g., flame pyrolysis) for 11% of samples, while the remaining 24% did not report the synthesis method. The median of primary size, exposure time, concentration, and surface area were 25 nm, 24 hour, 75 µg/mL, and 109 m^2/g, respectively. Abbreviations were provided in the dataset (Nanosilica Dataset File).

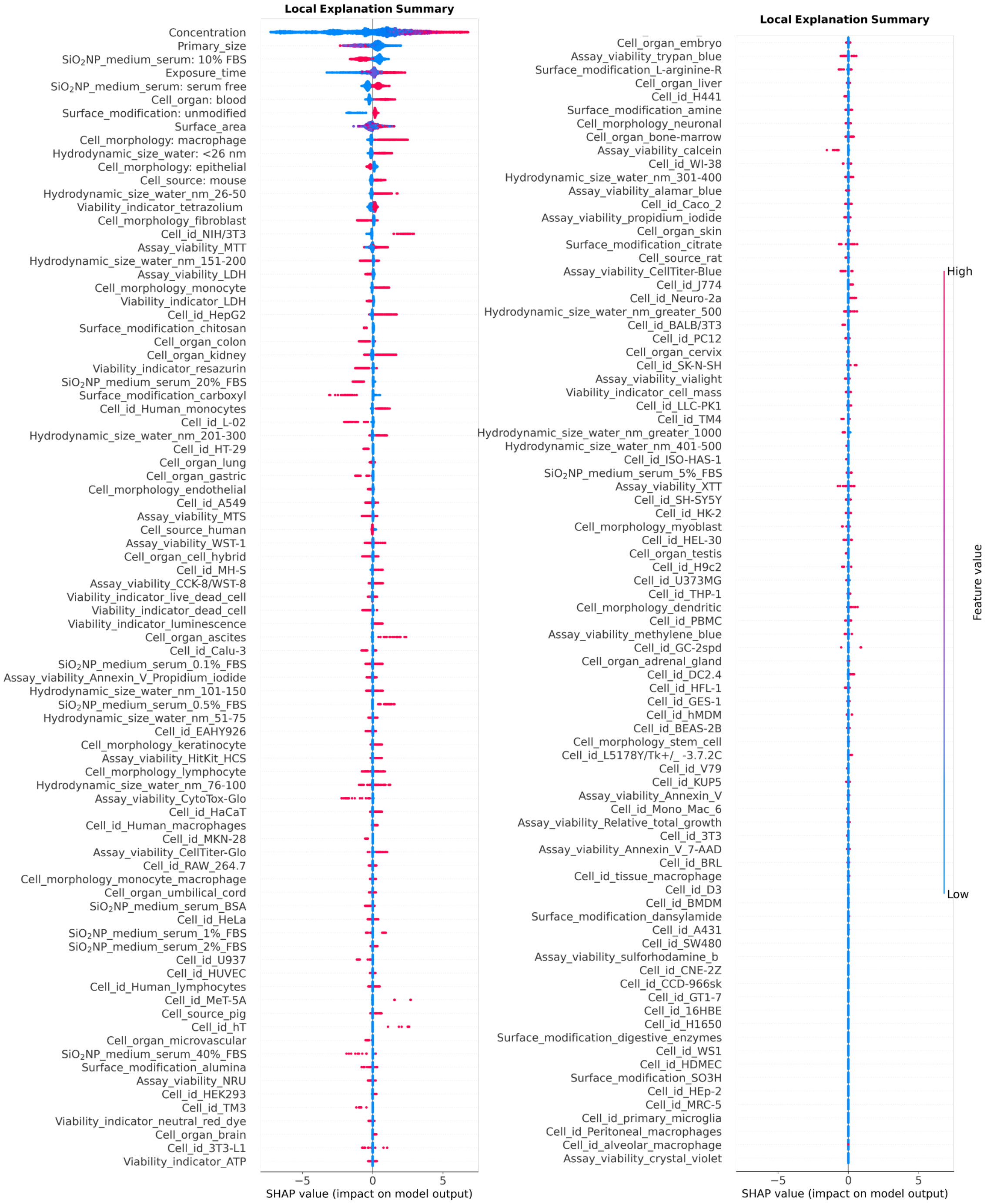


Figure S2. Complete Local Interpretability by CatBoost Model (Main Dataset: 4124 samples). For categorical attributes, red indicates the presence of attribute value, whereas blue indicates the absence of attribute value. For example, red of *Cell_id_NIH/3T3* means that the presence of NIH/3T3 cells drives the output of the model towards cytotoxicity, whereas blue drives the output towards noncytotoxicity.

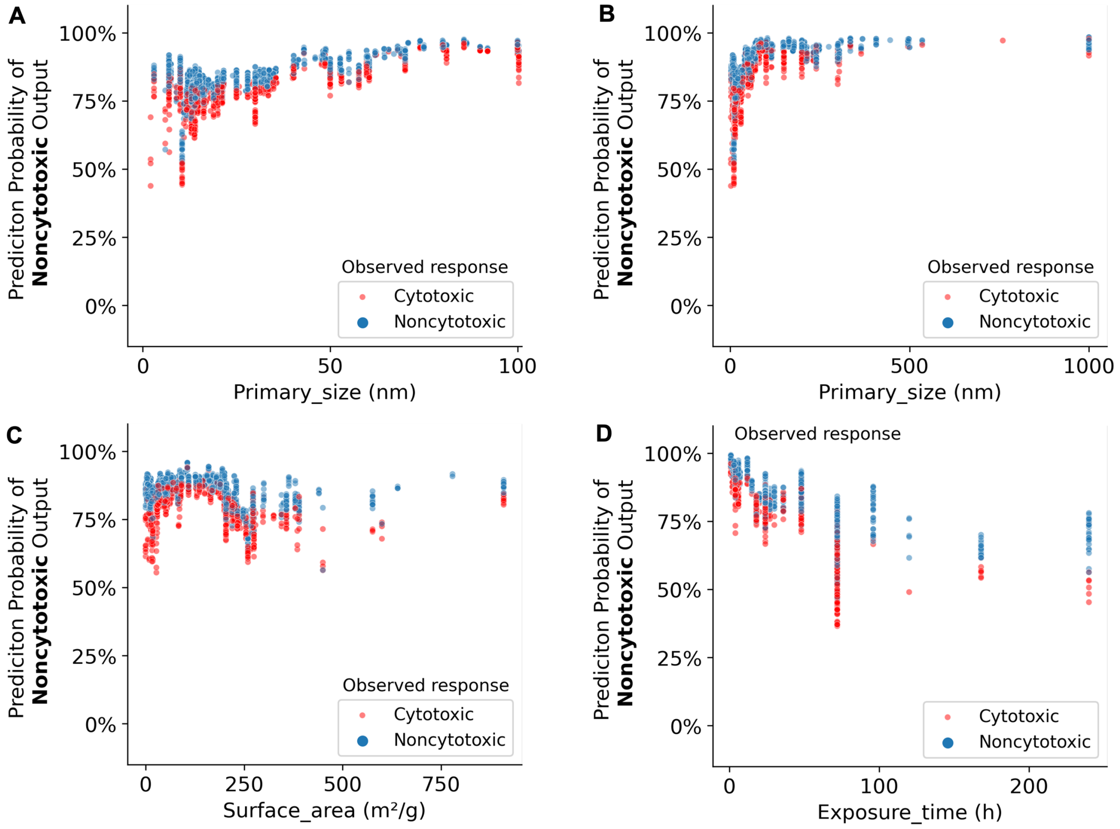


Figure S3. Prediction Probability of Noncytotoxity by CatBoost Model (Main Dataset: 4124 samples). The prediction probability of CatBoost to output noncytotoxic class at a given condition of one attribute only: (A,B) Primary size, (C) Surface area, and (D) Exposure time.

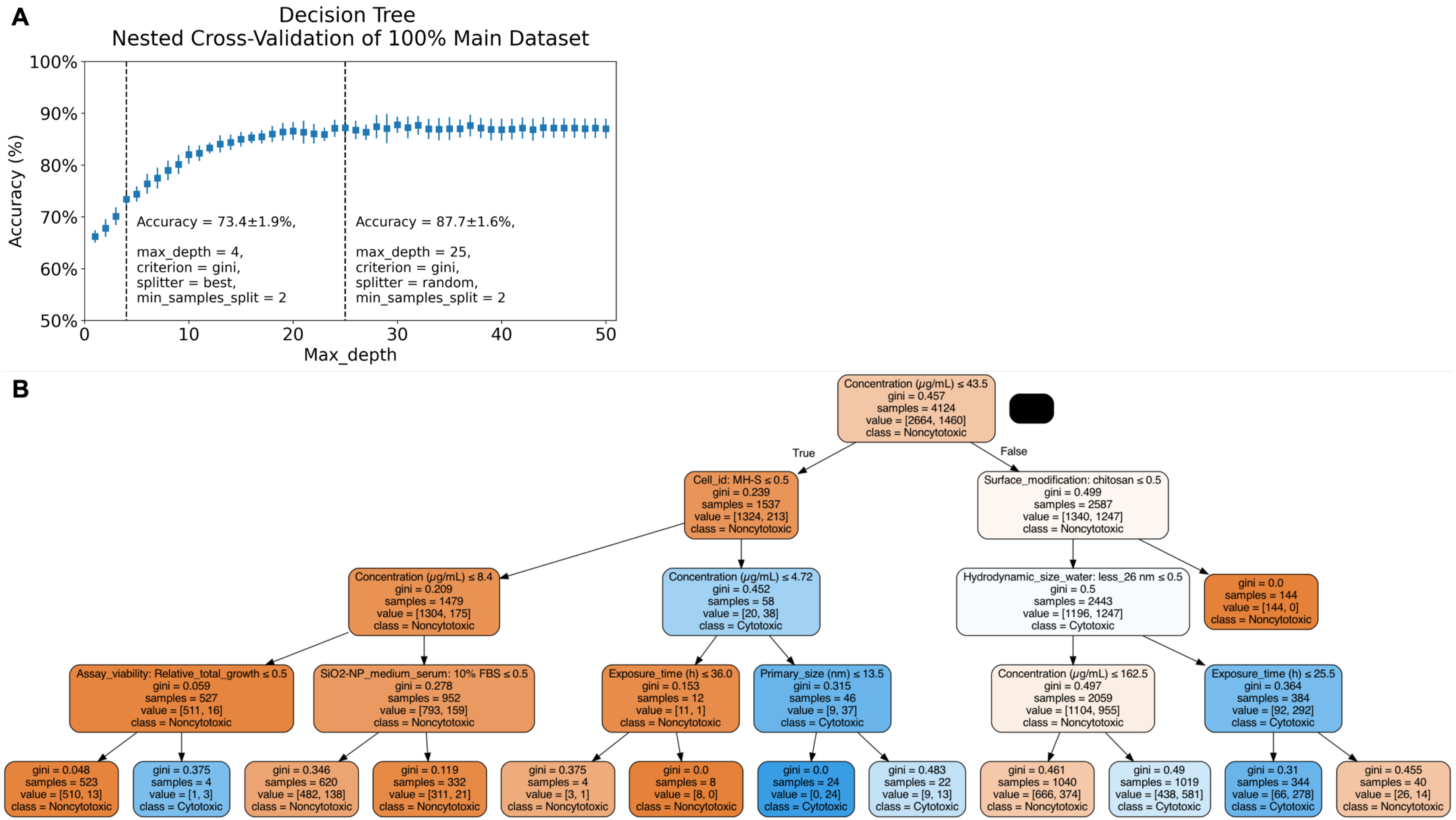

A
Decision Tree
Nested Cross-Validation of 100% Main Dataset
100%
90%
80%
70%
60%
50%
Accuracy (%)
0
10
20
30
40
50
Max_depth
Accuracy = 73.4±1.9%,
max_depth = 4,
criterion = gini,
splitter = best,
min_samples_split = 2
Accuracy = 87.7±1.6%,
max_depth = 25,
criterion = gini,
splitter = random,
min_samples_split = 2
B
Concentration (μg/mL) ≤ 43.5 gini = 0.457 samples = 4124 value = [2664, 1460] class = Noncytotoxic
[redacted]
True
False
Cell_id: MH-S ≤ 0.5 gini = 0.239 samples = 1537 value = [1324, 213] class = Noncytotoxic
Surface_modification: chitosan ≤ 0.5 gini = 0.499 samples = 2587 value = [1340, 1247] class = Noncytotoxic
Concentration (μg/mL) ≤ 8.4 gini = 0.209 samples = 1479 value = [1304, 175] class = Noncytotoxic
Concentration (μg/mL) ≤ 4.72 gini = 0.452 samples = 58 value = [20, 38] class = Cytotoxic
Hydrodynamic_size_water: less_26 nm ≤ 0.5 gini = 0.5 samples = 2443 value = [1196, 1247] class = Cytotoxic
gini = 0.0 samples = 144 value = [144, 0] class = Noncytotoxic
Assay_viability: Relative_total_growth ≤ 0.5 gini = 0.059 samples = 527 value = [511, 16] class = Noncytotoxic
SiO2-NP_medium_serum: 10% FBS ≤ 0.5 gini = 0.278 samples = 952 value = [793, 159] class = Noncytotoxic
Exposure_time (h) ≤ 36.0 gini = 0.153 samples = 12 value = [11, 1] class = Noncytotoxic
Primary_size (nm) ≤ 13.5 gini = 0.315 samples = 46 value = [9, 37] class = Cytotoxic
Concentration (μg/mL) ≤ 162.5 gini = 0.497 samples = 2059 value = [1104, 955] class = Noncytotoxic
Exposure_time (h) ≤ 25.5 gini = 0.364 samples = 384 value = [92, 292] class = Cytotoxic
gini = 0.048 samples = 523 value = [510, 13] class = Noncytotoxic
gini = 0.375 samples = 4 value = [1, 3] class = Cytotoxic
gini = 0.346 samples = 620 value = [482, 138] class = Noncytotoxic
gini = 0.119 samples = 332 value = [311, 21] class = Noncytotoxic
gini = 0.375 samples = 4 value = [3, 1] class = Noncytotoxic
gini = 0.0 samples = 8 value = [8, 0] class = Noncytotoxic
gini = 0.0 samples = 24 value = [0, 24] class = Cytotoxic
gini = 0.483 samples = 22 value = [9, 13] class = Cytotoxic
gini = 0.461 samples = 1040 value = [666, 374] class = Noncytotoxic
gini = 0.49 samples = 1019 value = [438, 581] class = Cytotoxic
gini = 0.31 samples = 344 value = [66, 278] class = Cytotoxic
gini = 0.455 samples = 40 value = [26, 14] class = Noncytotoxic

Figure S4. Decision Tree (Main Dataset: 4124 samples). (A) The nested cross-validation (10-fold) accuracy of decision tree vs. the depth of decision tree (max_depth). (B) A single decision tree with max_depth of 4 (nested cross-validation accuracy: 73.4±1.9%; accuracy of 64.6% $[\frac{2664}{1460+2664}]$ will be obtained by a model that always generates a noncytotoxic class). Notably, 86.1% $(\frac{1324}{213+1324})$ of silica nanoparticles with concentrations ≤43.5 µg/mL was associated with noncytotoxicity.

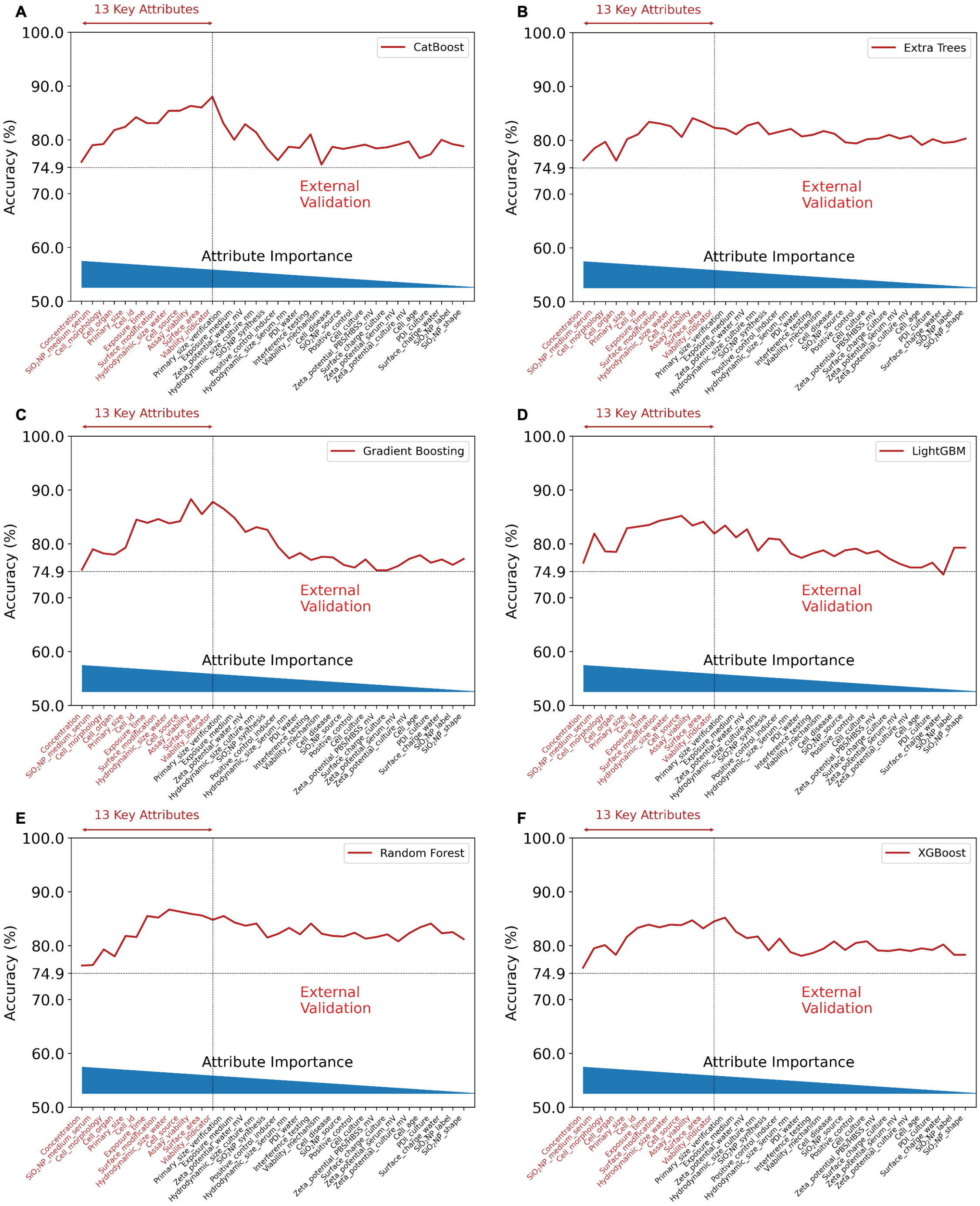


Figure S5. Predictive Accuracy of External Validation with Incrementally Added Attributes. An accuracy of 75.1% $(\frac{680}{680+225})$ will be attained by a model that always generates a noncytotoxic class in this independent dataset (905 samples).

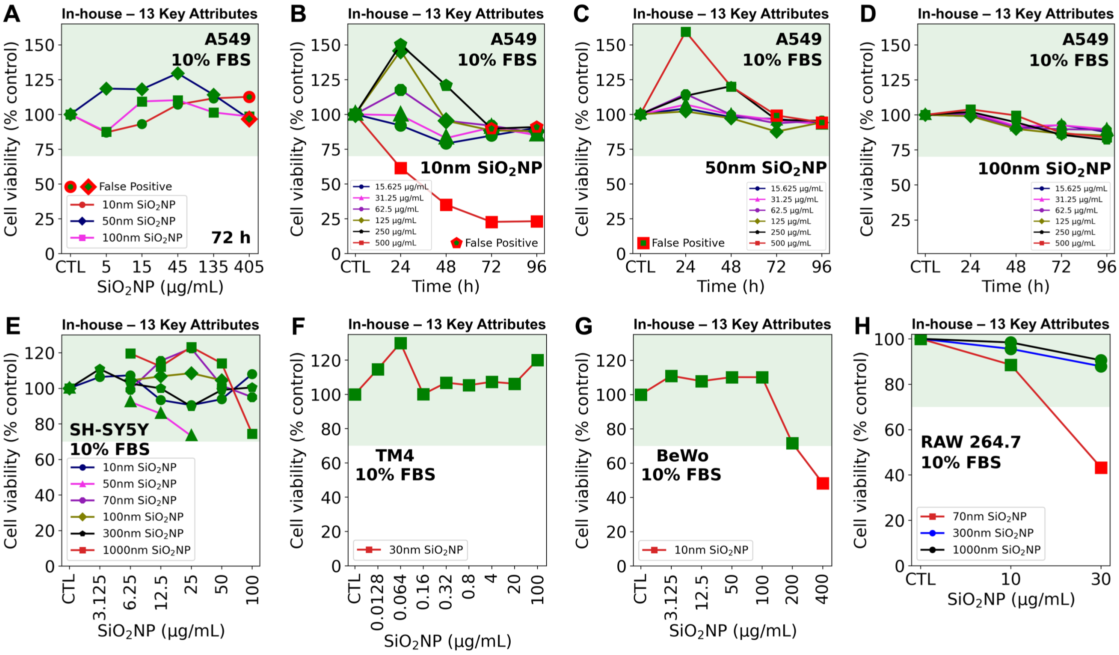
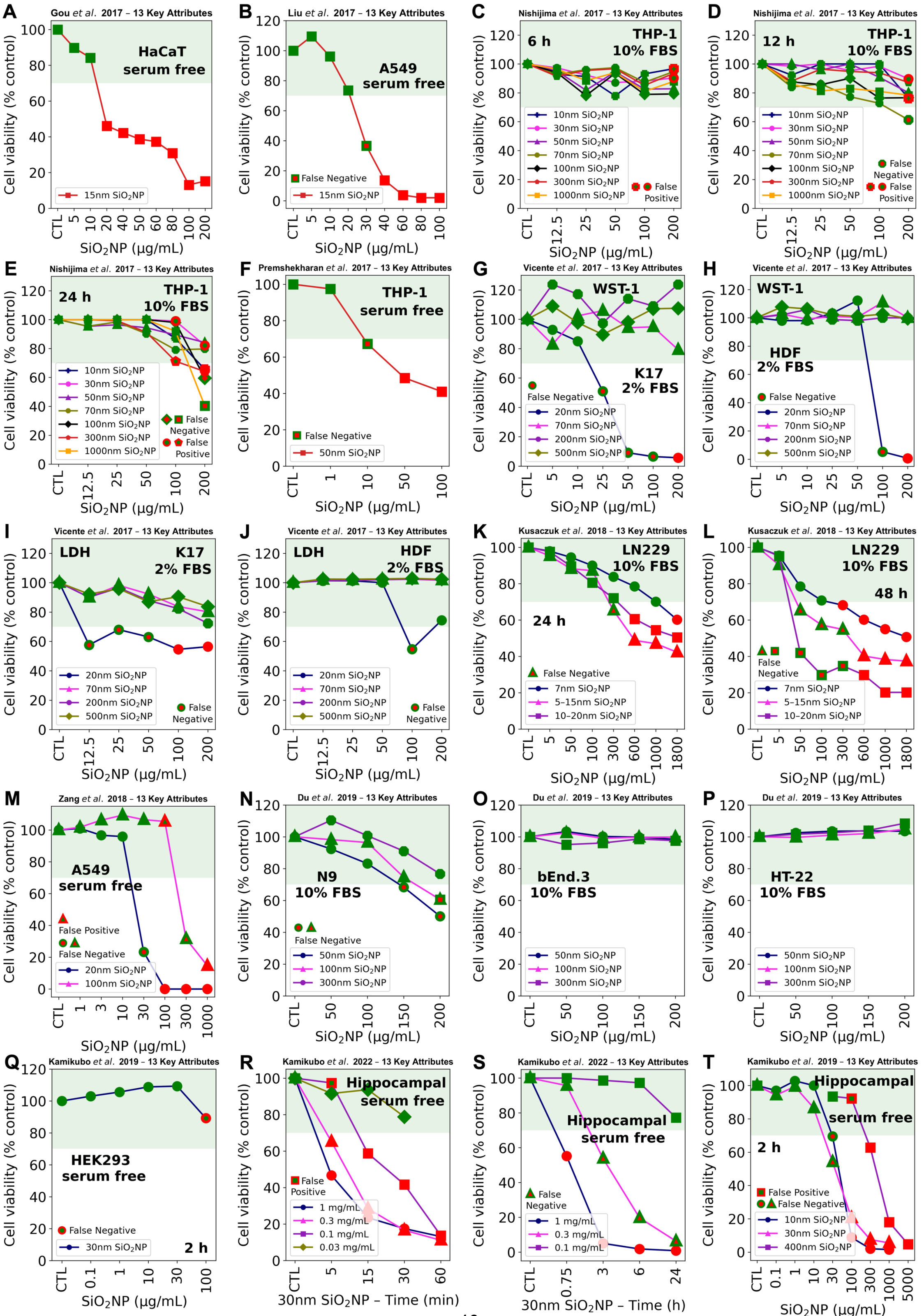
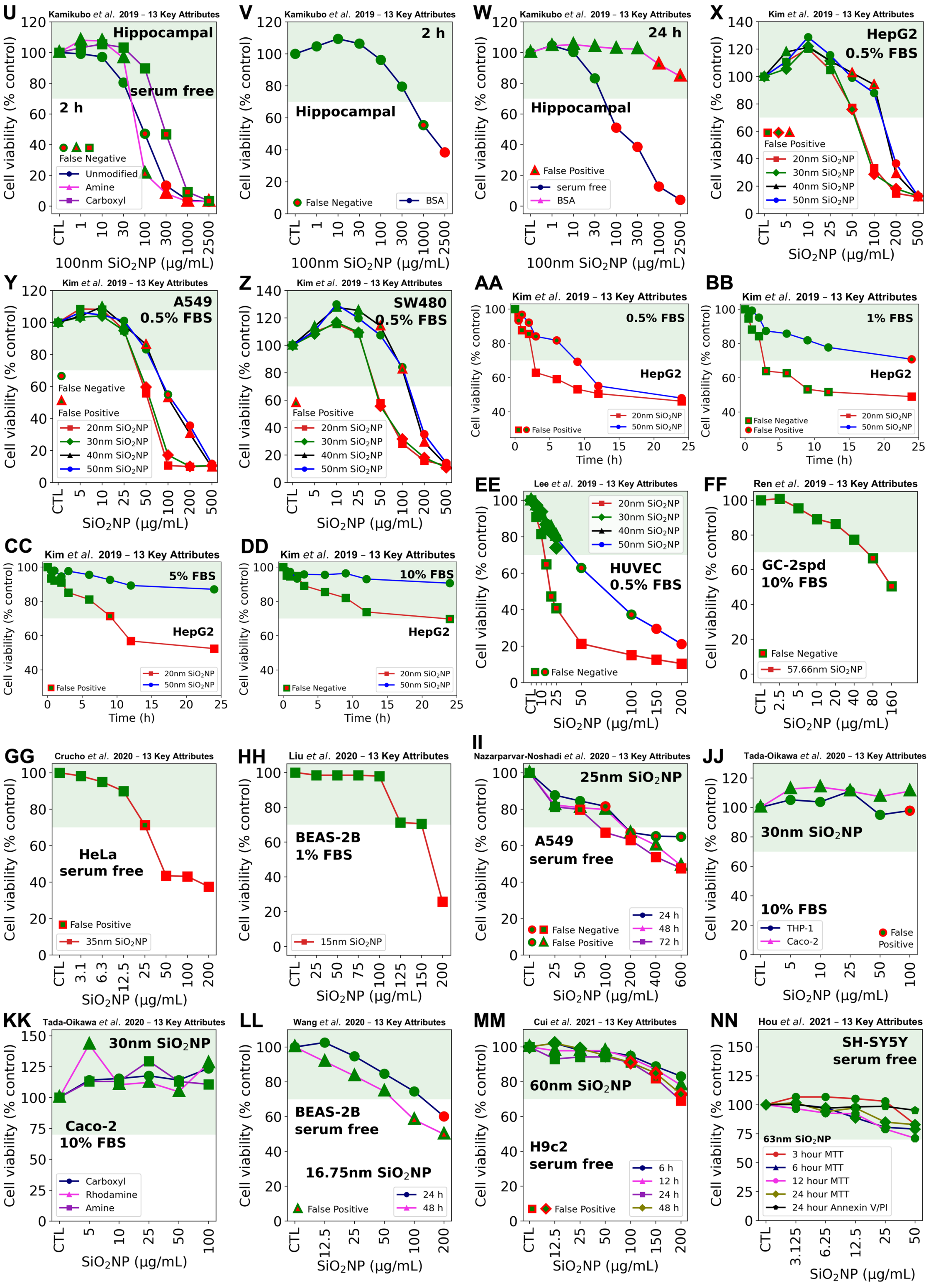


Figure S6. Prediction Errors of CatBoost Model for Eight Sets of In-house Experiments (136 Samples). Green and red markers indicate noncytotoxicity and cytotoxicity, respectively. Correctly classified samples have either a green or red marker, whereas misclassified samples have markers that are a combination of both colors.



A
Gou et al. 2017 – 13 Key Attributes
HaCaT
serum free
15nm SiO$_2$NP
B
Liu et al. 2017 – 13 Key Attributes
A549
serum free
False Negative
15nm SiO$_2$NP
C
Nishijima et al. 2017 – 13 Key Attributes
6 h
THP-1
10% FBS
10nm SiO$_2$NP
30nm SiO$_2$NP
50nm SiO$_2$NP
70nm SiO$_2$NP
100nm SiO$_2$NP
300nm SiO$_2$NP
1000nm SiO$_2$NP
False Positive
D
Nishijima et al. 2017 – 13 Key Attributes
12 h
THP-1
10% FBS
False Negative
False Positive
E
Nishijima et al. 2017 – 13 Key Attributes
24 h
THP-1
10% FBS
False Negative
False Positive
F
Premshekharan et al. 2017 – 13 Key Attributes
THP-1
serum free
False Negative
50nm SiO$_2$NP
G
Vicente et al. 2017 – 13 Key Attributes
WST-1
K17
2% FBS
False Negative
20nm SiO$_2$NP
70nm SiO$_2$NP
200nm SiO$_2$NP
500nm SiO$_2$NP
H
Vicente et al. 2017 – 13 Key Attributes
WST-1
HDF
2% FBS
False Negative
I
Vicente et al. 2017 – 13 Key Attributes
LDH
K17
2% FBS
False Negative
J
Vicente et al. 2017 – 13 Key Attributes
LDH
HDF
2% FBS
False Negative
K
Kusaczuk et al. 2018 – 13 Key Attributes
LN229
10% FBS
24 h
False Negative
7nm SiO$_2$NP
5-15nm SiO$_2$NP
10-20nm SiO$_2$NP
L
Kusaczuk et al. 2018 – 13 Key Attributes
LN229
10% FBS
48 h
False Negative
M
Zang et al. 2018 – 13 Key Attributes
A549
serum free
False Positive
False Negative
20nm SiO$_2$NP
100nm SiO$_2$NP
N
Du et al. 2019 – 13 Key Attributes
N9
10% FBS
False Negative
50nm SiO$_2$NP
100nm SiO$_2$NP
300nm SiO$_2$NP
O
Du et al. 2019 – 13 Key Attributes
bEnd.3
10% FBS
P
Du et al. 2019 – 13 Key Attributes
HT-22
10% FBS
Q
Kamikubo et al. 2019 – 13 Key Attributes
HEK293
serum free
False Negative
30nm SiO$_2$NP
2 h
R
Kamikubo et al. 2022 – 13 Key Attributes
Hippocampal
serum free
False Positive
1 mg/mL
0.3 mg/mL
0.1 mg/mL
0.03 mg/mL
30nm SiO$_2$NP – Time (min)
S
Kamikubo et al. 2022 – 13 Key Attributes
Hippocampal
serum free
False Negative
1 mg/mL
0.3 mg/mL
0.1 mg/mL
30nm SiO$_2$NP – Time (h)
T
Kamikubo et al. 2019 – 13 Key Attributes
Hippocampal
serum free
2 h
False Positive
False Negative
10nm SiO$_2$NP
30nm SiO$_2$NP
400nm SiO$_2$NP
Cell viability (% control)
SiO$_2$NP (μg/mL)
CTL



U
Kamikubo et al. 2019 – 13 Key Attributes
Hippocampal
serum free
2 h
False Negative
Unmodified
Amine
Carboxyl
Cell viability (% control)
0
20
40
60
80
100
120
CTL
1
10
30
100
300
1000
2500
100nm SiO$_2$NP (µg/mL)
V
Kamikubo et al. 2019 – 13 Key Attributes
2 h
Hippocampal
False Negative
BSA
W
Kamikubo et al. 2019 – 13 Key Attributes
24 h
Hippocampal
False Positive
serum free
BSA
X
Kim et al. 2019 – 13 Key Attributes
HepG2
0.5% FBS
False Positive
20nm SiO$_2$NP
30nm SiO$_2$NP
40nm SiO$_2$NP
50nm SiO$_2$NP
CTL
5
10
25
50
100
200
500
SiO$_2$NP (µg/mL)
Y
Kim et al. 2019 – 13 Key Attributes
A549
0.5% FBS
False Negative
False Positive
Z
Kim et al. 2019 – 13 Key Attributes
SW480
0.5% FBS
False Positive
AA
Kim et al. 2019 – 13 Key Attributes
0.5% FBS
HepG2
False Positive
20nm SiO$_2$NP
50nm SiO$_2$NP
Time (h)
0
5
10
15
20
25
BB
Kim et al. 2019 – 13 Key Attributes
1% FBS
HepG2
False Negative
False Positive
CC
Kim et al. 2019 – 13 Key Attributes
5% FBS
HepG2
False Positive
DD
Kim et al. 2019 – 13 Key Attributes
10% FBS
HepG2
False Negative
EE
Lee et al. 2019 – 13 Key Attributes
HUVEC
0.5% FBS
False Negative
CTL
5
10
25
50
100
150
200
FF
Ren et al. 2019 – 13 Key Attributes
GC-2spd
10% FBS
False Negative
57.66nm SiO$_2$NP
CTL
2.5
5
10
20
40
80
160
GG
Crucho et al. 2020 – 13 Key Attributes
HeLa
serum free
False Positive
35nm SiO$_2$NP
CTL
3.1
6.3
12.5
25
50
100
200
HH
Liu et al. 2020 – 13 Key Attributes
BEAS-2B
1% FBS
15nm SiO$_2$NP
CTL
25
50
75
100
125
150
200
II
Nazarparvar-Noshadi et al. 2020 – 13 Key Attributes
25nm SiO$_2$NP
A549
serum free
False Negative
False Positive
24 h
48 h
72 h
CTL
25
50
100
200
400
600
JJ
Tada-Oikawa et al. 2020 – 13 Key Attributes
30nm SiO$_2$NP
10% FBS
THP-1
Caco-2
False Positive
CTL
5
10
25
50
100
KK
Tada-Oikawa et al. 2020 – 13 Key Attributes
30nm SiO$_2$NP
Caco-2
10% FBS
Carboxyl
Rhodamine
Amine
150
125
100
75
50
25
0
LL
Wang et al. 2020 – 13 Key Attributes
BEAS-2B
serum free
16.75nm SiO$_2$NP
False Positive
24 h
48 h
CTL
12.5
25
50
100
200
MM
Cui et al. 2021 – 13 Key Attributes
60nm SiO$_2$NP
H9c2
serum free
False Positive
6 h
12 h
24 h
48 h
CTL
12.5
25
50
100
150
200
NN
Hou et al. 2021 – 13 Key Attributes
SH-SY5Y
serum free
63nm SiO$_2$NP
3 hour MTT
6 hour MTT
12 hour MTT
24 hour MTT
24 hour Annexin V/PI
CTL
3.125
6.25
12.5
25
50

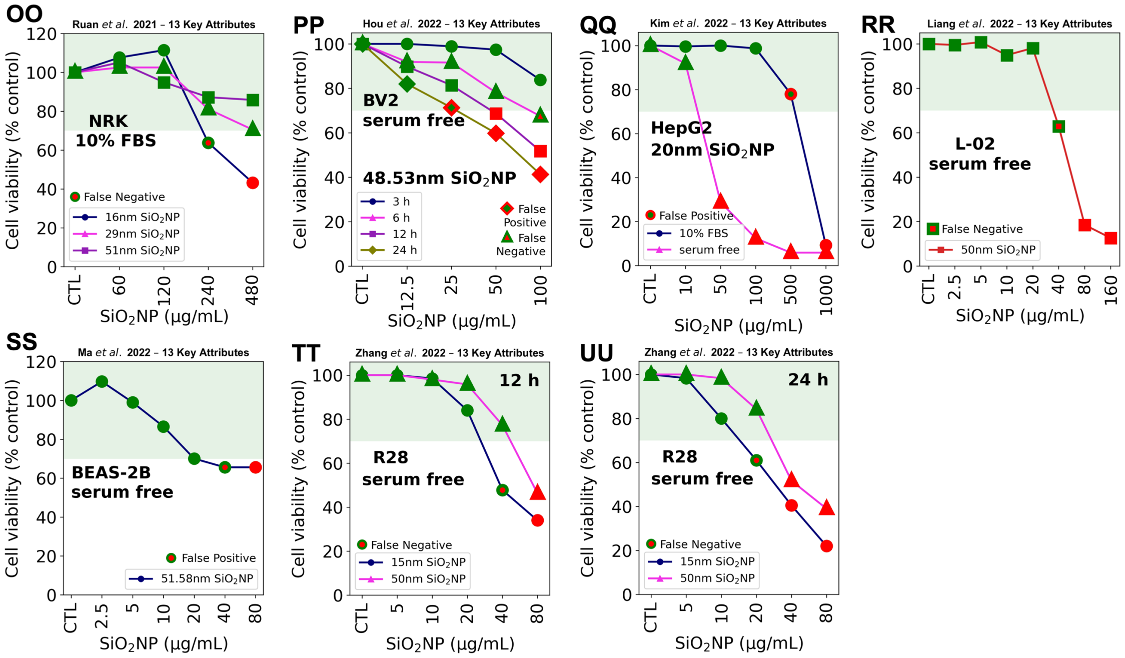


Figure S7. Prediction Errors of CatBoost Model for 47 Sets of Experiments (769 Samples). Green and red markers indicate noncytotoxicity and cytotoxicity, respectively. Correctly classified samples have either a green or red marker, whereas misclassified samples have markers that are a combination of both colors.

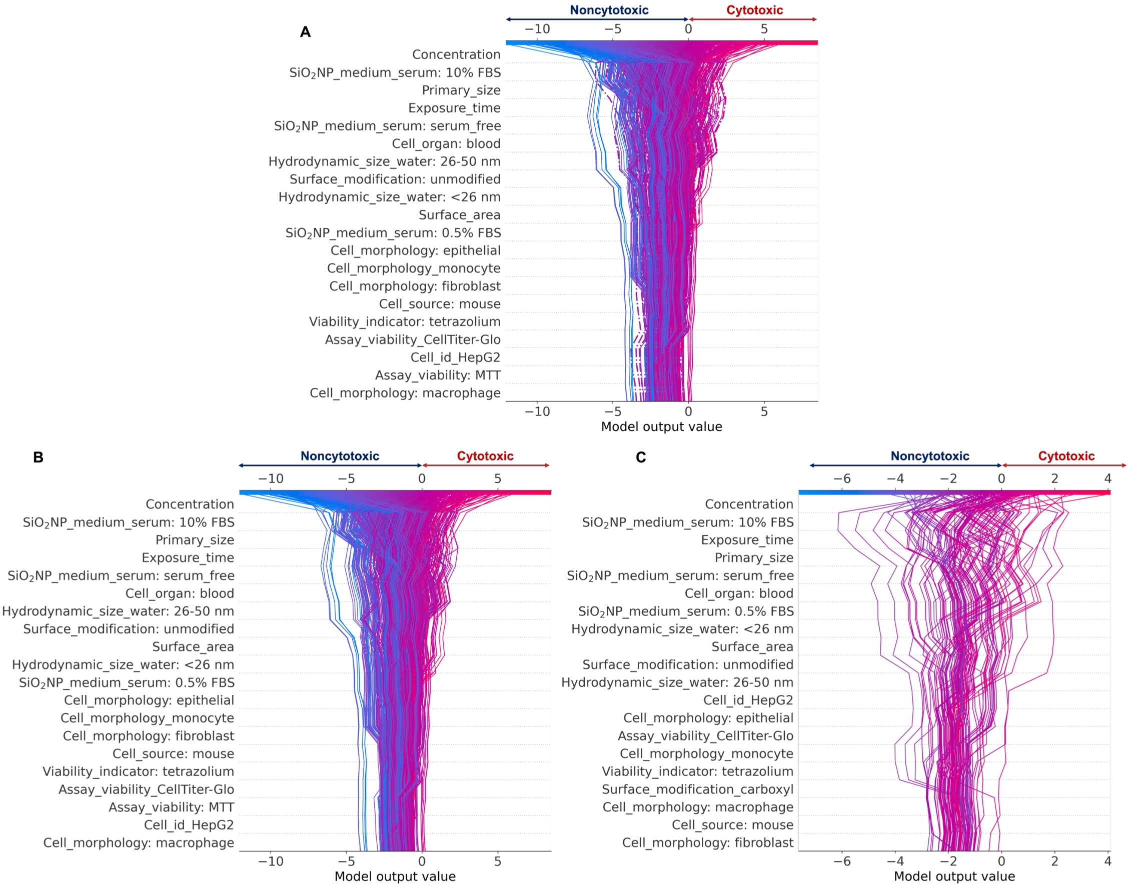


Figure S8. Decision Plots of Independent Dataset by CatBoost Model (905 samples). (A) Decision plot of predicted SiO_2-NP toxicity. Solid and dashed lines indicate correctly classified and misclassified samples, respectively. Separate decision plots of (B) correctly classified and (C) misclassified samples.

Table S1. List of Removed Dummy Features

No	Attributes	Removed Dummy Features
1	'Primary_size_verification',	'Primary_size_verification_not_verified',
2	'Hydrodynamic_size_water_nm’,	'Hydrodynamic_size_water_nm_not_determined',
3	'Hydrodynamic_size_culture_nm’,	'Hydrodynamic_size_culture_nm_not_determined',
4	'Hydrodynamic_size_serum_nm,	'Hydrodynamic_size_serum_nm_not_determined',
5	'PDI_water,	'PDI_water_not_determined',
6	'PDI_culture,	'PDI_culture_not_determined',
7	'Exposure_medium',	'Exposure_medium_M199',
8	'Positive_control’,	'Positive_control_not_included',
9	'Positive_control_inducer',	'Positive_control_inducer_not_available',
10	'Interference_testing ',	'Interference_testing_not_performed',
11	'SiO$_{2}$NP_medium_serum',	' SiO$_{2}$NP_medium_serum_15%_FBS',
12	'Zeta_potential_water_mV',	'Zeta_potential_water_mV_not_determined',
13	'Zeta_potential_PBS/HBSS_mV',	'Zeta_potential_PBS/HBSS_mV_not_determined',
14	'Zeta_potential_culture_mV ',	'Zeta_potential_culture_mV_not_determined',
15	'Zeta_potential_serum_mV,	'Zeta_potential_serum_mV_not_determined',
16	'Surface_charge_water,	'Surface_charge_water_positive',
17	'Surface_charge_culture,	'Surface_charge_culture_not_determined',
18	'Surface_modification,	'Surface_modification_CHO',
19	'SiO$_{2}$NP_label,	'SiO$_{2}$NP_label_none',
20	'SiO$_{2}$NP_source,	'SiO$_{2}$NP_source_in_house',
21	'SiO$_{2}$NP_synthesis,	'SiO$_{2}$NP_synthesis_not_available',
22	'SiO$_{2}$NP_shape,	'SiO$_{2}$NP_shape_rod',
23	'Cell_source,	'Cell_source_hamster',
24	'Cell_age,	'Cell_age_embryonic',
25	'Cell_id',	'Cell_id_MPMC/3t3',
26	'Cell_disease,	'Cell_disease_carcinoma',
27	'Cell_culture,	'Cell_culture_primary',
28	'Cell_organ,	'Cell_organ_heart',
29	'Cell_morphology,	'Cell_morphology_microglia',
30	'Assay_viability,	'Assay_viability_Sytox_Red',
31	'Viability_mechanism,	'Viability_mechanism_Cell_growth',
32	'Viability_indicator'	'Viability_indicator_live_cell'

Table S2. Hyperparameter Settings for 12 Machine Learning Algorithms for Internal Validation (80% Main Dataset)

Machine Learning	Hyperparameters (tested range)	Optimal Hyperparameters	
Linear		36 Attributes	13 Key Attributes
LDA	solver ∈ [‘svd’, ‘lsqr’, ‘eigen’]	solver = ‘lsqr’	solver = ‘lsqr’
LR	C ∈ [10^{-5}, 10^{-4}, 10^{-3}, ..., 10^{4}, 10^{5}]	C = 10^{5}	C = 10^{-2}
Ridge	alpha ∈ [10^{-5}, 10^{-4}, 10^{-3}, ..., 10^{4}, 10^{5}]	alpha = 10^{2}	alpha = 10^{2}
Nonlinear			
Kernel			
KNN	n_neighbors ∈ [1, 3, 5, ..., 97, 99]	n_neighbors = 1	n_neighbors = 1
SVM	C ∈ [1.0, 10^{2}, 10^{3}, ..., 10^{9}, 10^{10}], gamma ∈ [1.0, 10^{-1}, 10^{-2}, ..., 10^{-9}, 10^{-10}]	C = 10^{8}, gamma = 1	C = 10^{8}, gamma = 1
Tree-based			
DT	criterion ∈ [‘gini’,’entropy’], splitter ∈ ['best', 'random'], max_depth ∈ [20, 25, …, 40, 45, 50], min_samples_split ∈ [2, 3, 4]	criterion = entropy, splitter = random, max_depth = 35, min_samples_split = 2	criterion = gini, splitter = random, max_depth = 25, min_samples_split = 2
Extra Trees	n_estimators ∈ [100, 200, 300, …, 1400, 1500]	n_estimators = 800	n_estimators = 300
RF	n_estimators ∈ [100, 200, 300, …, 1400, 1500]	n_estimators = 300	n_estimators = 500

CatBoost	learning_rate ∈ [0.03, 0.04, 0.05, …, 0.09, 0.1], max_depth ∈ [3, 4, 5, …, 9, 10]	learning_rate = 0.07
		max_depth = 9
			learning_rate = 0.08
			max_depth = 7

GB	learning_rate ∈ [0.03, 0.04, 0.05, …, 0.09, 0.1], max_depth ∈ [3, 4, 5, …, 9, 10]	learning_rate = 0.09, max_depth = 7	learning_rate = 0.1, max_depth = 9
LightGBM	learning_rate ∈ [0.03, 0.04, 0.05, …, 0.09, 0.1], max_depth ∈ [3, 4, 5, …, 9, 10]	learning_rate = 0.09, max_depth = 10	learning_rate = 0.1, max_depth = 8
XGBoost	learning_rate ∈ [0.03, 0.04, 0.05, …, 0.09, 0.1], max_depth ∈ [3, 4, 5, …, 9, 10]	learning_rate = 0.09, max_depth = 10	learning_rate = 0.1, max_depth = 9

Footnotes: 10-fold nested cross validation with a fixed random state of 2022 was used in grid-search for inner evaluations (other hyperparameters were set to default values, feature scaling was applied for the linear and kernel classifiers). LDA, linear discriminant analysis; LR, logistic regression; DNN, deep neural network; KNN, k-nearest neighbors; SVM, support vector machine; DT, decision tree; Extra Trees, extremely randomized trees; RF, random forest; CatBoost, categorical boosting; GB, gradient boosting; LightGBM, light gradient boosting machine; XGBoost, extreme gradient boosting.

Table S3. Hyperparameter Settings for 12 Machine Learning Algorithms for External Validation (100% Main Dataset)

Machine Learning	Hyperparameters (tested range)	Optimal Hyperparameters	
Linear		13 Key Attributes	12 Key Attributes (dropped serum)
LDA	solver ∈ [‘svd’, ‘lsqr’, ‘eigen’]	solver = ‘lsqr’	solver = ‘lsqr’
LR	C ∈ [10^{-5}, 10^{-4}, 10^{-3}, ..., 10^{4}, 10^{5}]	C = 10^{4}	C = 10^{5}
Ridge	alpha ∈ [10^{-5}, 10^{-4}, 10^{-3}, ..., 10^{4}, 10^{5}]	alpha = 100	alpha = 0.1
Nonlinear			
Kernel			
KNN	n_neighbors ∈ [1, 3, 5, ..., 97, 99]	n_neighbors = 1	n_neighbors = 1
SVM	C ∈ [1.0, 10^{2}, 10^{3}, ..., 10^{9}, 10^{10}], gamma ∈ [1.0, 10^{-1}, 10^{-2}, ..., 10^{-9}, 10^{-10}]	C = 10^{8}, gamma = 0.1	C = 10^{7}, gamma = 1
Tree-based			
DT	criterion ∈ [‘gini’,’entropy’], splitter ∈ ['best', 'random'], max_depth ∈ [20, 25, …, 40, 45, 50], min_samples_split ∈ [2, 3, 4]	criterion = gini, splitter = random, max_depth = 30, min_samples_split = 2	criterion = entropy, splitter = random, max_depth = 50, min_samples_split = 3
Extra Trees	n_estimators ∈ [100, 200, 300, …, 1400, 1500]	n_estimators = 200	n_estimators = 700
RF	n_estimators ∈ [100, 200, 300, …, 1400, 1500]	n_estimators = 1500	n_estimators = 900

CatBoost	learning_rate ∈ [0.03, 0.04, 0.05, …, 0.09, 0.1], max_depth ∈ [3, 4, 5, …, 9, 10]	learning_rate = 0.05
		max_depth = 7
			learning_rate = 0.04
			max_depth = 7

GB	learning_rate ∈ [0.03, 0.04, 0.05, …, 0.09, 0.1], max_depth ∈ [3, 4, 5, …, 9, 10]	learning_rate = 0.09,
		max_depth = 8
			learning_rate = 0.08
			max_depth = 8

LightGBM	learning_rate ∈ [0.03, 0.04, 0.05, …, 0.09, 0.1], max_depth ∈ [3, 4, 5, …, 9, 10]	learning_rate = 0.09,
		max_depth = 9
			learning_rate = 0.1
			max_depth = 10

XGBoost	learning_rate ∈ [0.03, 0.04, 0.05, …, 0.09, 0.1], max_depth ∈ [3, 4, 5, …, 9, 10]	learning_rate = 0.09,
		max_depth = 10
			learning_rate = 0.1
			max_depth = 9

Footnotes: 10-fold nested cross validation with a fixed random state of 2022 was used in grid-search for inner evaluations (other hyperparameters were set to default values, feature scaling was applied for the linear and kernel classifiers). LDA, linear discriminant analysis; LR, logistic regression; DNN, deep neural network; KNN, k-nearest neighbors; SVM, support vector machine; DT, decision tree; Extra Trees, extremely randomized trees; RF, random forest; CatBoost, categorical boosting; GB, gradient boosting; LightGBM, light gradient boosting machine; XGBoost, extreme gradient boosting.

Table S4. Cell Viability Assay

A549 cell lines
Our in-house experiments used SiO_2-NPs with primary sizes of 10, 50, and 100 nm on human lung epithelial A549 cells for 24, 48, 72, and 96 h, with a total of 87 samples. A549 cells were obtained from American Type Culture Collection (ATCC; Manassas, VA, USA) and cultured in Dulbecco's modified Eagle's medium (DMEM, high glucose [4.5 g/L]) supplemented with 10% inactivated fetal bovine serum (FBS) and a 1% antibiotic cocktail at 37 °C in 5% CO_2. A549 cells were seeded in 96-well plates (10,000 cells/100 µL) and incubated overnight in culture media. Next, they were exposed to SiO_2-NPs for 24–96 h. Various concentrations of SiO_2-NPs (5–500 µg/mL) were prepared *via* dilution with 10% FBS-DMEM. Cell viability was evaluated using the WST-8 assay.
SH-SY5Y cell lines
Our in-house experiments also examined the effects of SiO_2-NPs with primary sizes of 10, 50, 70, 100, 300, and 1000 nm on the human-derived neuroblastoma cell line SH-SY5Y (29 samples). SH-SY5Y cells were acquired from ATCC and cultured in DMEM/Ham's F-12 supplemented with 10% inactivated FBS and a 1% antibiotic cocktail at 37 °C in 5% CO_2. SH-SY5Y cells were seeded in 96-well plates (20,000 cells/100 µL) and incubated overnight in culture media. They were then exposed to SiO_2-NPs for 72 h. SiO_2-NPs were diluted to various concentrations (3.125–100 µg/mL) in culture media containing serum. Again, WST-8 assay was used to examined cell viability.
TM4 cell lines
In in-house experiments, we employed SiO_2-NPs with primary sizes of 30 nm on mouse testis epithelial TM4 cells (eight samples). TM4 cells were obtained from ATCC and cultured in DMEM supplemented with 10% inactivated FBS and a 1% antibiotic cocktail at 37 °C in 5% CO_2. TM4 cells were seeded in 96-well plates (10,000 cells/50 µL) and incubated overnight in culture media before exposure to SiO_2-NPs for 24 h. SiO_2-NPs were prepared in various concentrations (0.0128–100 µg/mL) *via* dilution with culture media containing serum. Cell viability was also examined using the WST-8 assay.
BeWo cell lines
Our in-house experiments subjected human choriocarcinoma cell line BeWo (six samples) to SiO2-NPs with primary sizes of 10 nm. BeWo cells were acquired from the Japanese Collection of Research Bioresources Cell Bank (JCRB9111; Osaka, Japan) and cultured in Ham's F-12 medium supplemented with 10% inactivated FBS and a 1% antibiotic cocktail at 37 °C in 5% CO_2. BeWo cells were seeded in 96-well plates (1,000 cells/200 µL) and incubated overnight in culture media. Then, they were exposed to SiO_2-NPs for 48 h. Various concentrations of SiO_2-NPs (3.125–400 µg/mL) were prepared *via* dilution with culture media containing serum. Cell viability was measured with the colorimetric dye 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyl tetrazolium bromide (MTT, Tokyo Chemical Industry, Tokyo, Japan) in accordance with the manufacturer's instructions.
RAW 264.7 cell lines
In these in-house experiments, we used SiO_2-NPs with primary sizes of 70, 300, and 1000 nm on mouse macrophage RAW 264.7 cells (six samples). RAW 264.7 cells were obtained from ATCC and cultured in DMEM supplemented with 10% inactivated FBS and a 1% antibiotic cocktail at 37 °C in 5% CO_2. RAW 264.7 cells (1,500 cells) were seeded and then incubated overnight in culture media before being exposed to SiO_2-NPs for 120 h. SiO_2-NPs were diluted to 10 and 30 µg/mL in culture media containing serum. Cell viability was again evaluated using the WST-8 assay.

Table S5. Prediction-Error Comparisons: External Validation (12 Key Attributes and 905 Samples)

Machine Learning	$nCV_{10\text{-fold}}$	Accuracy	AUC-ROC	Recall	Precision
Linear					
LDA	71.6±1.6%	56.9%	62.4%	50.2%	28.9%
LR	73.1±1.9%	55.5%	62.1%	55.6%	29.2%
Ridge	72.9±1.7%	58.1%	65%	60.4%	31.9%
Nonlinear					
DNN	73.3±2.2%	63.9%	63.5%	52.8%	35.1%
Kernel					
KNN	84.1±1.0%	68.1%	64.8%	58.2%	40.2%
SVM	84.7±1.9%	76.1%	50%	4.0%	100.0%
Tree-based					
DT	84.7±0.8%	72.6%	63.0%	46.7%	45.1%
Extra Trees	85.3±1.2%	79.4%	81.0%	44.0%	62.3%
RF	86.5±1.3%	79.6%	83.3%	43.1%	63.0%
CatBoost	88.7±1.3%	80.7%	84.4%	53.3%	63.2%
GB	88.1±1.6%	81.3%	86.1%	62.2%	62.5%
LightGBM	87.2±1.6%	78.7%	83.0%	56.4%	57.2%
XGBoost	87.1±1.3%	79.9%	84.1%	54.7%	60.6%

Footnotes: The predictive models with 12 key attributes were built by dropping the *SiO_2-NP_medium_serum* attribute. LDA, linear discriminant analysis; LR, logistic regression; DNN, deep neural network; KNN, k-nearest neighbors; SVM, support vector machine; DT, decision tree; Extra Trees, extremely randomized trees; RF, random forest; CatBoost, categorical boosting; GB, gradient boosting; LightGBM, light gradient boosting machine; XGBoost, extreme gradient boosting.

Chapter 2: Computational drug design guided by machine learning for urate transporter 1 (URAT1)

Background

In renal tubules, urate transporter 1 (URAT1) facilitates the reabsorption of over 90% of uric acid through an ion-exchange mechanism. Given its significant role, URAT1 emerges as a promising target for the development of innovative anti-hyperuricemic drugs. Currently, five major uricosuric drugs—probenecid, sulfinpyrazone, benzbromarone, lesinurad, and dotinurad—have entered the market, and verinurad, a lesinurad analogue, is undergoing phase II clinical studies. However, the clinical use of existing URAT1 inhibitors is constrained by severe adverse effects, notably liver and kidney toxicity, necessitating the development of safer and more potent URAT1 inhibitors.[72–74]

Computer-aided drug design can boost drug development efficiency and involves two primary methods: structure-based drug design and ligand-based drug design (LBDD). In the absence of a high-resolution three-dimensional (3D) structure of URAT1, LBDD emerges as a promising approach for drug discovery, relying solely on small molecule information. A crucial consideration for LBDD is the utilization of available data resources to identify new and innovative leads. While many researchers commonly employ ChEMBL[75] as a key resource for activity data, it is important to note that ChEMBL curates a limited dataset specific to a target protein rather than drawing from global evidence.

In this study, we present an LBDD pipeline aimed at identifying promising potential lead compounds for URAT1 by uniting literature data mining and machine learning. Due to its training on the physicochemical properties and fingerprints of compounds, the proposed LBDD pipeline has the potential to proficiently pinpoint unique compound skeletons with similar traits. The pipeline initiates with the extraction of URAT1 inhibitors data from a comprehensive dataset sourced from global evidence (aggregate of scientific and patent publications). Subsequently, a prediction model is constructed to distinguish between high and low active inhibitors, employing key descriptors and a counteractivity explanation. The model is then utilized to generate innovative potential lead compounds from a massive ZINC database, incorporating the model's probability, principal component, Tanimoto coefficient, and predicted absorption, distribution, metabolism, and excretion (ADME) parameters by DruMAP[76].

Methods

Literature data mining

Three comprehensive patent reviews on the activity of URAT1 inhibitors served as the basis for the literature data evaluation in this study. These reviews, conducted by Pan *et al.*,[72] Dong *et al.* ,[73] and

Shi *et al.*[74] covered patent literature published up to 2015, 2019, and 2023 respectively. The PubMed database was also systematically searched for scientific publications using the following search strategy: (uric acid transporter 1 [tiab] OR uric acid transporter [tiab] OR urate transporter 1 [tiab] OR urate transporter [tiab] OR URAT1 [tiab] OR hURAT1 [tiab] OR urate reabsorption [tiab] OR potent uric acid [tiab] OR antihyperuricemic [tiab] OR antihyperuricemia [tiab]) AND (discovery [tiab] OR design [tiab] OR synthesis [tiab] OR structure–activity relationship [tiab] OR structure–activity relationships [tiab]OR SAR [tiab] OR derivatives [tiab] OR derivative [tiab] OR ligand [tiab] OR analog [tiab] OR analogs [tiab] analogue [tiab] OR analogues [tiab] OR ligands [tiab] OR compounds [tiab] OR compound [tiab] OR scaffolds [tiab] OR scaffold [tiab] OR novel [tiab]). The selection of literature adhered to the PICOS framework[32] of evidence-based medicine, ensuring consistency and reliability across the chosen studies. The criteria for inclusion were: (1) a population involving URAT1; (2) the intervention and comparison focusing on test compounds vs. a positive control; (3) the outcome centered on IC_{50} (half-maximal inhibitory concentration) as the activity metric; and (4) the study design being an *in vitro* structure-activity relationship study. Exclusion criteria encompassed non-isolated compounds (*e.g.*, extracts), abstract articles, and other non-relevant studies. In total, 25 scientific[77–101] and 75 patent[72–74] publications meeting the inclusion criteria were incorporated.

Systematic extraction of structural information with SMILES (simplified molecular-input line-entry system) and IC_{50} resulted in a dataset comprising 2717 nonduplicate compounds. De-duplication of compounds was implemented by excluding the minority class and subsequently computing the mean of IC_{50} values. For example, if there are seven IC_{50} values, where five belong to high activity and two to low activity, the mean was calculated using the five IC_{50} values from the high active class. The high active class has IC_{50} values less than or equal 500 nM, and the low active class has IC_{50} value greater than 500 nM. A threshold of 500 nM was implemented based on IC_{50} values of major uricosuric inhibitors exhibiting low (lesinurad [IC_{50} = 6.5 μM],[84] probenecid [IC_{50} = 15 μM],[84] sulfinpyrazone [IC_{50} = 716 μM][80]) and high activity (benzbromarone [IC_{50} = 280 nM],[89] verinurad [IC_{50} = 170 nM],[100] and dotinurad [IC_{50} = 360 nM, patent CN112430221B]). The selected threshold aimed to identify compounds with IC_{50} values similar to highly potent clinical trial compounds (benzbromarone, verinurad, dotinurad). It also led to a balanced dataset, with the ratio of high and low active URAT1 inhibitors approaching 1:1. Maintaining a balanced dataset is crucial to prevent machine learning model from exhibiting bias towards a specific class. IC_{50} were converted to binary labels: "1" (≤500 nM, high active) and "0" (>500 nM, low active). SMILES was standardized using ChEMBL_Structure_Pipeline[102] package.

Feature engineering

Physicochemical descriptors were calculated using Mordred (v1.2.0)[103], RDKit, and MOE

(Molecular Operating Environment, v2022.02) software. Molecular fingerprints were calculated using RDKit to describe MACCS (166 bits) and Morgan fingerprints (2048 bits with radius 2, 2048 bits with radius 3, 4096 bits with radius 2, and 4096 bits with radius 3). Molecular fingerprints serve as a technique for portraying a molecule through a series of binary bits, indicating either activation ("1") or deactivation ("0"), while still retaining essential information about the molecular composition. Physicochemical descriptors and molecular fingerprints were combined, yielding a total of 13,959 features.

Initially, feature selection involved eliminating features with Pearson correlation coefficients exceeding 0.9, resulting in 8,874 features. The Boruta package, acting as a wrapper for a Random Forest classification algorithm, was utilized to identify relevant features with a threshold set at 99.999999%, resulting in a selection of 186 features.

Machine learning

Nine established machine-learning algorithms were utilized (1) linear discriminant analysis (LDA), (2) logistic regression (LR), (3) k-nearest neighbors (KNN), (4) decision tree (DT), (5) random forest (RF), (6) categorical boosting (CatBoost), (7) gradient boosting (GB), (8) light gradient boosting machine (LightGBM), and (9) extreme gradient boosting (XGBoost), categorized as linear (1-2), nonlinear kernels (3), and nonlinear tree-based classifiers (4-9). The scikit-learn (v1.0.2), CatBoost (v1.0.4), LightGBM (v3.3.2), and XGBoost (v1.5.1) packages in Python 3.10 were employed for implementation.

The dataset, containing 2717 samples, underwent shuffling and was split into training (80%) and test (20%) sets *via* random stratified sampling. Predictive models with binary classification were initially developed using 80% of the dataset, and the 186 selected features were incrementally added to those previously selected by the Shapley Additive exPlanation (SHAP)[104], employing the CatBoost algorithms. Split-sample validation was conducted with the remaining 20% of the dataset, independent of model building, and SHAP values were used to identify key features that contributed to optimal predictive performance. Predictive models were then reconstructed using 80% of the dataset with the identified key features, using the 9 machine-learning algorithms. Final predictive models were constructed using the entire dataset and identified key features as a final training set. Ten-fold nested cross-validation (nCV)[105] was applied, and the models were fine-tuned using GridSearchCV to obtain optimal hyperparameters. SHAP values of the identified key features were calculated for principal component analysis. Counteractive explanations, implemented through the use of the exmol[106] package, were employed to address the question of "what is the smallest alteration to the compound structures that would modify their activity". Essentially, a counteractive compound closely resembles the original compound but results in a different activity.

The evaluation metrics were based on AUC-ROC (area under the curve of the true-positive rate or recall [$\frac{TP}{TP\ +\ FN}$] vs. false-positive rate [$\frac{FP}{FP+\ TN}$]), recall, and precision $\left(\frac{TP}{TP\ +\ FP}\right)$, and accuracy (1 − $\frac{TP\ +\ TN}{TP\ +\ FP+\ TN+FN}$), where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively.

Prioritization of potential lead compounds from ZINC database

ZINC15 database, available at https://zinc15.docking.org/, was obtained, encompassing 4,167,324 purchasable (in-stock) compounds with lead-like and fragment-like properties. Potential hit compounds were identified by considering (1) the optimal prediction probability from the final predictive model, (2) principal component 1 (PC1) generated by SHAP vectors of the identified key features using the final predictive model, and (3) structural similarity between ZINC compounds and the final training dataset. The maximum common substructure-based Tanimoto coefficient[107] scores on RDKit were employed to compute structural similarity, and a golden ratio of Tanimoto scores less than 0.382 was used to select novel potential hit scaffolds. Potential leads were generated by utilizing DruMAP to predict ADME of the novel potential hit scaffolds, with hepatic intrinsic clearance in liver microsome (CL_{int}), fraction absorbed (F_a), and fraction unbound in plasma ($f_{u,p}$) as ADME metrics.

Results

Literature data curation

We gathered URAT1 inhibitors data for 2717 compounds, sourcing from 100 publications spanning the years 2007 to 2023. Utilizing the 500 nM cutoff, 42% of the compounds were high active, while 58% were low active.

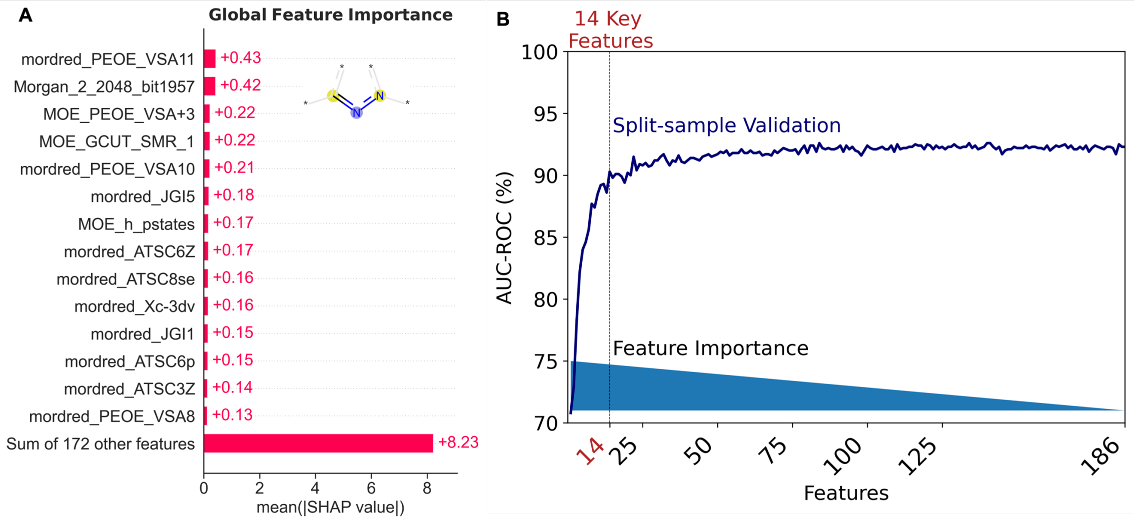
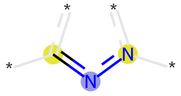


Figure 2.1. Feature importance for URAT1 inhibitors, based on CatBoost. (A) Global interpretability for the average absolute SHAP value magnitudes. (B) Predictive AUC-ROC of split-sample validation with incrementally increasing features. SHAP, Shapley Additive exPlanations.

Feature Importance

We utilized feature importance for identifying key feature *via* SHAP values with CatBoost. Based on the feature importance (Figure 2.1A), we identified the top 14 features that resulted to optimal predictive AUC-ROC (Figure 2.1B), arranging them in order of importance: *mordred_PEOE_VSA11,* *Morgan_2_2048_bit1957, MOE_PEOE_VSA+3, MOE_GCUT_SMR_1, mordred_PEOE_VSA10, mordred_JGI5, MOE_h_pstates, mordred_ATSC6Z, mordred_ATSC8se, mordred_Xc-3dv, mordred_JGI1, mordred_ATSC6p, mordred_ATSC3Z,* and *mordred_PEOE_VSA8.* The description of each feature is shown in Table 2.1.

Table 2.1. Description of features

Descriptor	Description
mordred_PEOE_VSA11	MOE Charge VSA Descriptor 11 (0.15 <= x < 0.20).
Morgan_2_2048_bit1957	
MOE_PEOE_VSA+3	Sum of vi where qi is in the range [0.15,0.20), where v_i be the van der Waals surface area (Å^2) of atom *i* (as calculated by a connection table approximation).
MOE_GCUT_SMR_1	The GCUT descriptors using atomic contribution to molar refractivity (using the Wildman and Crippen SMR method) instead of partial charge.
mordred_PEOE_VSA10	MOE Charge VSA Descriptor 10 (0.10 <= x < 0.15).
mordred_JGI5	5-ordered mean topological charge.
MOE_h_pstates	The entropic count or fractional number of protonation states.
mordred_ATSC6Z	Centered moreau-broto autocorrelation of lag 6 weighted by atomic number.
mordred_ATSC8se	Centered moreau-broto autocorrelation of lag 8 weighted by sanderson EN.
mordred_Xc-3dv	3-ordered Chi cluster weighted by valence electrons.
mordred_JGI1	1-ordered mean topological charge.
mordred_ATSC6p	Centered moreau-broto autocorrelation of lag 6 weighted by polarizability.
mordred_ATSC3Z	Centered moreau-broto autocorrelation of lag 3 weighted by atomic number.
mordred_PEOE_VSA8	MOE Charge VSA Descriptor 8 (0.00 <= x < 0.05).

Split-sample validation

A primary evaluation metric of the split-sample validation is AUC-ROC (Table 2.2). Tree-based classifiers demonstrated satisfactory AUC-ROCs ranging from 89.0% to 90.3%, with CatBoost exhibiting the best performance (AUC-ROC: 90.3%, precision: 82.1%, recall: 76.3%, and accuracy: 83.1%). Linear, KNN, and DT exhibited AUC-ROCs of 76.9% to 86.0%.

Table 2.2. Prediction-error comparisons: Split-sample validation (14 key features and 20% test set)

Machine Learning	AUC-ROC	Precision	Recall	Accuracy
Linear				
LDA	78.2%	75.3%	52.2%	72.8%
LR	78.3%	69.9%	53.9%	71.0%
Nonlinear				
Kernel				
KNN	86.0%	74.9%	73.2%	78.5%
Tree-based				
DT	76.9%	70.3%	68.4%	74.6%
RF	89.0%	82.4%	71.9%	81.8%
CatBoost	90.3%	82.1%	76.3%	83.1%
GB	89.0%	79.0%	74.1%	80.9%
LightGBM	90.0%	79.9%	75.0%	81.6%
XGBoost	89.0%	79.6%	73.7%	81.1%

Footnotes: LDA, linear discriminant analysis; LR, logistic regression; KNN, k-nearest neighbors; DT, decision tree; RF, random forest; CatBoost, categorical boosting; GB, gradient boosting; LightGBM, light gradient boosting machine; XGBoost, extreme gradient boosting.

Final predictive model

We constructed final predictive models using the 14 key features and entire dataset (2717 compounds) and obtained robust nCV AUC-ROCs (Table 2.3), with CatBoost exhibiting satisfactory performance (AUC-ROC: 89.3±1.3%, precision: 79.8±2.3%, recall: 77.7±3.8%, and accuracy: 82.4±1.7%).

Table 2.3. Prediction-error comparisons: 10-fold nCV (14 key features and entire dataset)

Machine Learning	AUC-ROC	Precision	Recall	Accuracy
Linear				
LDA	77.1±2.2%	71.1±3.6%	50.6±2.6%	70.7±1.7%
LR	77.2±1.9%	68.7±4.1%	53.7±2.9%	70.3±2.3%
Nonlinear				
Kernel				
KNN	86.2±2.6%	73.6±4.1%	76.4±3.9%	78.5±2.7%
Tree-based				
DT	78.0±3.3%	70.1±4.6%	68.6±5.1%	74.5±3.1%
RF	89.4±2.2%	79.7±4.3%	76.8±3.3%	82.0±2.7%
CatBoost	89.3±1.3%	79.8±2.3%	77.7±3.8%	82.4±1.7%
GB	88.4±2.0%	77.8±2.9%	74.9±2.8%	80.4±1.3%
LightGBM	88.5±2.5%	78.0±5.2%	76.6±4.6%	81.0±2.9%
XGBoost	88.6±2.2%	78.3±4.2%	76.6±3.7%	81.2±2.5%

Footnotes: LDA, linear discriminant analysis; LR, logistic regression; KNN, k-nearest neighbors; DT, decision tree; RF, random forest; CatBoost, categorical boosting; GB, gradient boosting; LightGBM, light gradient boosting machine; XGBoost, extreme gradient boosting.

We analyzed the SHAP-value distribution of the 14 key features across the final training set using CatBoost. According to the SHAP local explanation summary (Figure 2.2A), a larger value of *MOE_PEOE_VSA+3, mordred_PEOE_VSA10, mordred_JGI5, MOE_h_pstates, mordred_PEOE_VSA11,* and *mordred_ATSC3Z* were associated with reduced activity effects. In contrast, a higher value of *Morgan_2_2048_bit1957, MOE_GCUT_SMR_1, mordred_ATSC6Z, mordred_ATSC8se, mordred_Xc-3dv, mordred_JGI1, mordred_ATSC6p,* and *mordred_PEOE_VSA8* were associated with high activity.

A visualization plot was generated to compare the CatBoost model's predicted and observed pIC_{50} (negative logarithm of IC_{50}) values, revealing a correlation between the model's probability and precision performance (Figure 2.2B). Specifically, a precision of 95.9% can be attained when the model's probability output is ≥ 98% (log odds ≥ 3.9). Additionally, an analysis of the SHAP vector of 14 key features showed a correlation between PC1 (23.3% of explained SHAP vector variance) and precision performance (Figure 2.2C), where a precision of 95.2% is associated with PC1 values ≤ – 2.67.

We then proceeded to provide a quantitative analysis of URAT1 inhibitors, distinguishing between those with high and low activity, utilizing 14 key descriptor features with verinurad and lesinurad as representative inhibitors (Figure 2.3). We derived counteractive compounds to explore minimal alterations to the structures of both verinurad and lesinurad that could impact their respective activity levels. In essence, counteractive compounds closely resemble the originals but result in a shift from high to low activity in verinurad and vice versa in lesinurad. Through counteractive explanations, our analysis indicated that modifying the teal-colored carboxylic acid group or methyl group of verinurad could offer insights into the specific substructures responsible for its high activity (Figure 2.3A). Additionally, introducing an additional NH or Br group to lesinurad was identified as a potential enhancement for its activity (Figure 2.3B).

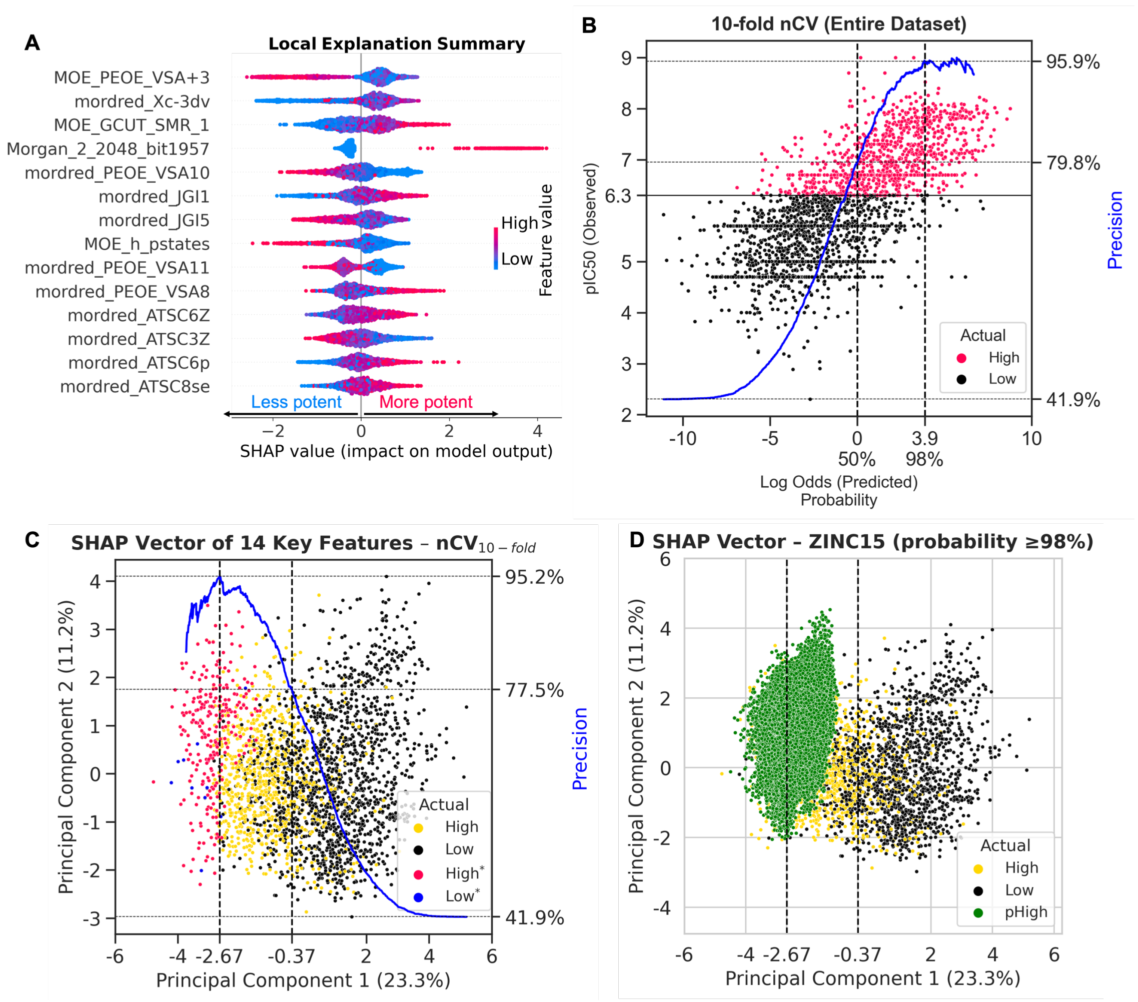


Figure 2.2. Final model, based on CatBoost. (A) Local interpretability, with each dot corresponding to a compound of URAT1 inhibitors obtained from entire dataset. (B) Prediction-error of nested cross validation using 14 key features from entire dataset. (C) SHAP vector using 14 key features from entire dataset. PC1 and PC2 explains 23.3% and 11.2% of SHAP vector variance, respectively. High*, compounds exhibiting ≥98% probability of belonging to high active class. (D) Screening of ZINC compounds using final CatBoost model. pHigh, ZINC compounds exhibiting ≥98% probability of belonging to high active class. nCV, nested cross-validation; PC, principal component; SHAP, Shapley Additive exPlanations; URAT1, urate transporter 1.

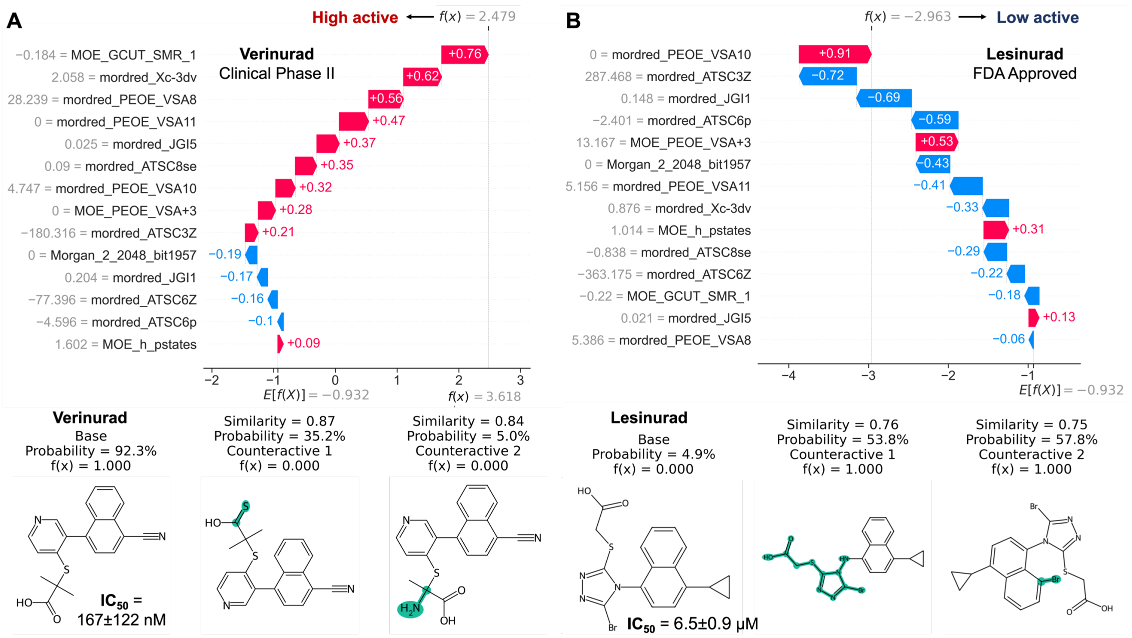


Figure 2.3. Activity explanation using physicochemical properties and molecular structures for (A) verinurad and (B) lesinurad, based on CatBoost. The positive and negative values of f(x) = 2.479 and f(x) = –2.963 correspond to the high active and low active classes, respectively. Teal color represents alterations made to the base molecule, while counteractive depicts the specific modifications that render the activity of base molecule against URAT1. URAT1, urate transporter 1

Prioritization of promising potential lead compounds from ZINC database

To generate potential leads from ZINC15 database, we implemented various filtering steps (Figure 2.4). Initially, we removed any duplicated ZINC compounds, yielding 3,457,766 compounds. We predicted this list of compounds using final CatBoost model and kept 42,594 compounds exhibiting ≥ 98% probability (corresponds to 95.9% precision). We then filter these compounds using PC1 and kept 9,760 compounds exhibiting PC1 values ≤ –2.67 (corresponds to 95.2% precision, Figure 2.2D). Next, we identified 7,082 novel potential hit compounds using Tanimoto coefficient. Finally, we obtained

22 promising potential lead compounds exhibiting good ADME properties, *i.e.,* stable CL_{int} of less than 20 µL/min/mg, high F_a of more than 0.7, and low $f_{u,p}$ of less than 0.05 using DruMAP.

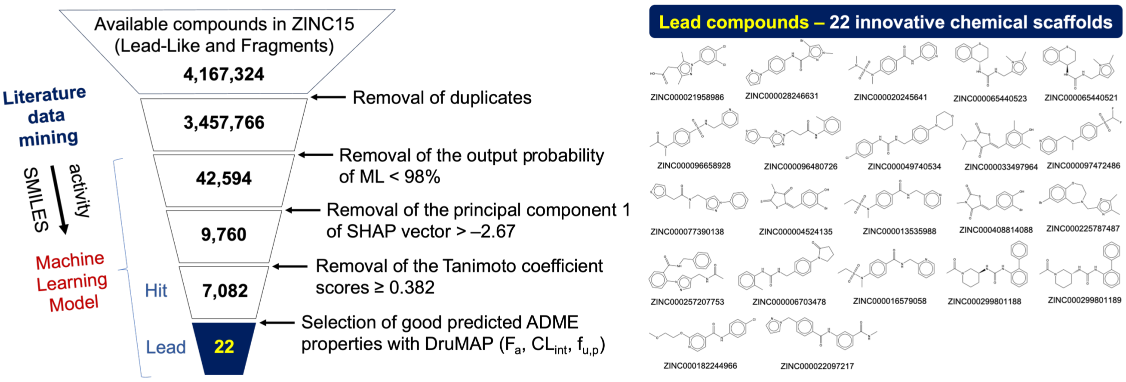


Figure 2.4. Ligand-based drug design (LBDD) pipeline for identifying innovative potential lead compounds from ZINC database. CL_{int}, hepatic intrinsic clearance in liver microsome; F_a, fraction absorbed; $f_{u,p}$, fraction unbound in plasma.

Discussion

The bespoke LBDD pipeline presents a promising workflow for identifying novel chemical scaffolds targeting URAT1 without a high-resolution 3D structure. The pipeline is currently confined to URAT1 but holds the potential to extend its applicability to other targets. To our knowledge, this study is the first LBDD approach incorporating global evidence to develop machine learning models to discriminate high from low active URAT1 inhibitors. Other methods like transfer learning could provide alternative strategy, by using a large chemical space as the training source. Compared with a dataset from ChEMBL (approximately 600 compounds for URAT1 or SLC22A12), we have generated the largest dataset, containing 2717 URAT1 inhibitors, to train a successful machine learning model. Moreover, our CatBoost model offers insights into the different class activities of URAT1 inhibitors by utilizing only 14 key features (13 physicochemical properties and 1 fingerprint) and counteractively generated compounds in the analysis. Other tree-based algorithms were not explored to identify key features due to computational constraints, considering CatBoost had already shown satisfactory performance. The focus was on the practical utility of the model, and the choice of CatBoost was driven by its effectiveness and efficiency for our specific objectives.

An important part of our analysis involves attaining 95% precision, assessed from two perspectives: the model's probability and principal component of SHAP values. By supplementing these with the use of Tanimoto coefficient and DruMAP, we have identified 22 promising commercially available potential lead compounds with distinct skeletons, showing a similarity below 32.8% when compared with the known compounds. All these compounds are predicted to have ideal properties: high solubility (>10 μg/mL) at pH 7.4, stable intrinsic clearance (< 20 μL/min/mg), high fraction of a dose absorbed (>0.7), high apparent permeability coefficient (>100 nm/s), and low fraction unbound in plasma (<0.05). We anticipate that these compounds undergo experimental validation successfully and demonstrate potential therapeutic efficacy with favorable pharmacokinetic profiles against hyperuricemia.

Apart from employing machine learning models for screening the ZINC database to acquire novel skeletons, scaffold hopping serves as a method to obtain compounds with distinct skeletons, and it is not strictly dependent on machine learning techniques. The process of scaffold hopping can be executed using chemical intuition.[108]

A potential direction for future research involves developing a unified platform that encompasses activity, kinetics, and toxicity predictions, aiming to enhance the success rate of drug development. While incorporating DruMAP for predicted ADME assessment in the current LBDD pipeline is a positive step, leveraging other publicly available web servers, such as ProTox-II,[109] could offer supplementary information on the predicted toxicities of small molecules.

Acknowledgments

I would like to express sincere gratitude to Professor Kenji Mizuguchi, Associate Professor Kosuke Hashimoto, Assistant Professor Reiko Watanabe, and Lecturer Gert-Jan Bekker for their great support in carrying out this research.

I would like to thank Professor Yasuo Tsutsumi, Associate Professor Kazuma Higashisaka, and Assistant Professor Yuya Haga for their in-house experiments data.

I would also like to thank everyone in this laboratory for their advice and cooperation during the research.

I gratefully acknowledge the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) for the study grant at Osaka University.

Finally, I would like to thank my family and friends for their continued support.

Publications, Conferences, Books, and Presentations in PhD

Publications

1. **Martin***, Watanabe, R., Hashimoto, K., Higashisaka, K., Haga, Y., Tsutsumi, Y., Mizuguchi, K.* Evidence-Based Prediction of Cellular Toxicity for Amorphous Silica Nanoparticles. *ACS Nano* **2023**, 17, 9987–9999, DOI: 10.1021/acsnano.2c11968
2. **Martin***, Zhou, Y., Takagi, T., Tian, YS.* Safety, efficacy, and cost-effectiveness of insulin degludec U100 versus insulin glargine U300 in adults with type 1 diabetes: a systematic review and indirect treatment comparison. *Int J Clin Pharm* **2022**, *44*, 587–598, DOI: 10.1007/s11096-022-01410-x
3. **Martin**, Zhou, Y., Takagi, T., Tian, YS.* Efficacy and safety among second-generation and other basal insulins in adult patients with type 1 diabetes: a systematic review and network meta-analysis. *Naunyn-Schmiedeberg's Arch Pharmacol* **2021**, *394*, 2091–2101, DOI: 10.1007/s00210-021-02128-9

Conferences

1. **Martin***, *et al.* Prediction of Xanthine Oxidase Inhibitors using Graph Neural Network. *Chem-Bio Informatics Society (CBI) Annual Meeting 2023*, Tokyo, Japan, October 23–26, 2023 (Poster)
2. **Martin***, *et al.* Prediction of Monoacylglycerol Lipase (MAGL) Inhibitors using Explainable Artificial Intelligence. *2023 年日本バイオインフォマティクス学会年会第12回生命医薬情報学連合大会（IIBMP2023）*, Chiba, Japan, September 7–9, 2023 (Poster)

Books

1. **Martin**・渡邉怜子・水口賢司. ナノ粒子をより安全に 設計するための新手法──ナノ粒子の安全性に革命を： AI が拓く未来. *化学同人*. 化学 2023 年 11 月号

Presentations

1. **Martin** ・ Computational Drug Design Guided by Machine Learning for Urate Transporter 1 (URAT1). Presented at the Institute for Protein Research Retreat (IPR Retreat), November 2023.

* Corresponding author

References

1. Oh, E. *et al.* Meta-analysis of cellular toxicity for cadmium-containing quantum dots. *Nat Nanotechnol* **11**, 479–486 (2016).
2. Bilal, M. *et al.* Bayesian Network Resource for Meta-Analysis: Cellular Toxicity of Quantum Dots. *Small* **15**, e1900510 (2019).
3. Gernand, J. M. & Casman, E. A. A meta-analysis of carbon nanotube pulmonary toxicity studies-how physical dimensions and impurities affect the toxicity of carbon nanotubes. *Risk Analysis* **34**, 583–597 (2014).
4. Ma, Y., Wang, J., Wu, J., Tong, C. & Zhang, T. Meta-analysis of cellular toxicity for graphene via data-mining the literature and machine learning. *Science of the Total Environment* **793**, 148532 (2021).
5. Xu, J., Lin, X. & Gowen, A. A. Combining machine learning with meta-analysis for predicting cytotoxicity of micro- and nanoplastics. *Journal of Hazardous Materials Advances* **8**, 100175 (2022).
6. Labouta, H. I., Asgarian, N., Rinker, K. & Cramb, D. T. Meta-Analysis of Nanoparticle Cytotoxicity via Data-Mining the Literature. *ACS Nano* **13**, 1583–1594 (2019).
7. Liu, L. *et al.* Cytotoxicity of phytosynthesized silver nanoparticles: A meta-analysis by machine learning algorithms. *Sustain Chem Pharm* **21**, 100425 (2021).
8. Kad, A., Pundir, A., Arya, S. K., Puri, S. & Khatri, M. Meta-analysis of in-vitro cytotoxicity evaluation studies of zinc oxide nanoparticles: Paving way for safer innovations. *Toxicology in Vitro* **83**, 105418 (2022).

9. Puzyn, T. *et al.* Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. *Nat Nanotechnol* **6**, 175–178 (2011).

10. Manganelli, S., Leone, C., Toropov, A. A., Toropova, A. P. & Benfenati, E. QSAR model for predicting cell viability of human embryonic kidney cells exposed to SiO2 nanoparticles. *Chemosphere* **144**, 995–1001 (2016).

11. Choi, J. S., Ha, M. K., Trinh, T. X., Yoon, T. H. & Byun, H. G. Towards a generalized toxicity prediction model for oxide nanomaterials using integrated data from different sources. *Sci Rep* **8**, 6110 (2018).

12. Fortino, V. *et al.* Biomarkers of nanomaterials hazard from multi-layer data. *Nat Commun* **13**, 3798 (2022).

13. Walkey, C. D. *et al.* Protein corona fingerprinting predicts the cellular interaction of gold and silver nanoparticles. *ACS Nano* **8**, 2439–2455 (2014).

14. Chen, R. *et al.* Nanoparticle surface characterization and clustering through concentration-dependent surface adsorption modeling. *ACS Nano* **8**, 9446–9456 (2014).

15. Ho, S. Y., Phua, K., Wong, L. & Bin Goh, W. W. Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability. *Patterns* **1**, 100129 (2020).

16. Steckler, A. & McLeroy, K. R. The importance of external validity. *Am J Public Health* **98**, 9–10 (2008).

17. Bleeker, S. E. *et al.* External validation is necessary in prediction research: A clinical example. *J Clin Epidemiol* **56**, 826–832 (2003).

18. Monopoli, M. P., Åberg, C., Salvati, A. & Dawson, K. A. Biomolecular coronas provide the biological identity of nanosized materials. *Nat Nanotechnol* **7**, 779–786 (2012).

19. Fytianos, G., Rahdar, A. & Kyzas, G. Z. Nanomaterials in cosmetics: Recent updates. *Nanomaterials* **10**, 979 (2020).

20. Mebert, A. M., Baglole, C. J., Desimone, M. F. & Maysinger, D. Nanoengineered silica: Properties, applications and toxicity. *Food and Chemical Toxicology* **109**, 753–770 (2017).

21. Miyata, T. *et al.* Nanoscale Stress Distribution in Silica-Nanoparticle-Filled Rubber as Observed by Transmission Electron Microscopy: Implications for Tire Application. *ACS Appl Nano Mater* **4**, 4452–4461 (2021).

22. Mizutani, T., Arai, K., Miyamoto, M. & Kimura, Y. Application of silica-containing nano-composite emulsion to wall paint: A new environmentally safe paint of high performance. *Prog Org Coat* **55**, 276–283 (2006).

23. Bernauer, U. *et al.* The SCCS scientific advice on the safety of nanomaterials in cosmetics. *Regulatory Toxicology and Pharmacology* **126**, 105046 (2021).

24. Croissant, J. G., Butler, K. S., Zink, J. I. & Brinker, C. J. Synthetic amorphous silica nanoparticles: toxicity, biomedical and environmental implications. *Nat Rev Mater* **5**, 886–909 (2020).

25. Murugadoss, S. *et al.* Toxicology of silica nanoparticles: an update. *Arch Toxicol* **91**, 2967–3010 (2017).

26. Napierska, D., Thomassen, L. C. J., Lison, D., Martens, J. A. & Hoet, P. H. The nanosilica hazard: Another variable entity. *Part Fibre Toxicol* **7**, 39 (2010).

27. Biological evaluation of medical devices – Part 5: Tests for in vitro cytotoxicity. ISO 10993-5. Available at: https://www.iso.org/obp/ui/#iso:std:iso:10993:-5:en. (2009).

28. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on

Medical Devices, Amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and Repealing Council Directives 90/385/EEC and 93/42/EE.

29. Dong, X. *et al.* The Size-dependent Cytotoxicity of Amorphous Silica Nanoparticles: A Systematic Review of in vitro Studies. *Int J Nanomedicine* **15**, 9089–9113 (2020).

30. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* **2**, 56–67 (2020).

31. Cawley, G. C. & Talbot, N. L. C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research* **11**, 2079–2107 (2010).

32. Schardt, C., Adams, M. B., Owens, T., Keitz, S. & Fontelo, P. Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Med Inform Decis Mak* **7**, 16 (2007).

33. Marin, F., Rohatgi, A. & Charlot, S. WebPlotDigitizer, a polyvalent and free software to extract spectra from old astronomical publications: application to ultraviolet spectropolarimetry. *arXiv[Preprint]* (2017).

34. Gong, C., Yang, L., Zhou, J., Guo, X. & Zhuang, Z. Possible role of PAPR-1 in protecting human HaCaT cells against cytotoxicity of SiO2 nanoparticles. *Toxicol Lett* **280**, 213–221 (2017).

35. Liu, W. *et al.* Nrf2 protects against oxidative stress induced by SiO2 nanoparticles. *Nanomedicine* **12**, 2303–2318 (2017).

36. Nishijima, N. *et al.* Human scavenger receptor A1-mediated inflammatory response to silica particle exposure is size specific. *Front Immunol* **8**, 1–12 (2017).

37. Premshekharan, G., Nguyen, K., Zhang, H., Forman, H. J. & Leppert, V. J. Low dose inflammatory potential of silica particles in human-derived THP-1 macrophage cell culture studies – Mechanism and effects of particle size and iron. *Chem Biol Interact* **272**, 160–171

(2017).

38. Vicente, S., Moia, C., Zhu, H. & Vigé, X. In vitro evaluation of the internalization and toxicological profile of silica nanoparticles and submicroparticles for the design of dermal drug delivery strategies. *Journal of Applied Toxicology* **37**, 1396–1407 (2017).

39. Kusaczuk, M., Krętowski, R., Naumowicz, M., Stypułkowska, A. & Cechowska-Pasko, M. Silica nanoparticle-induced oxidative stress and mitochondrial damage is followed by activation of intrinsic apoptosis pathway in glioblastoma cells. *Int J Nanomedicine* **13**, 2279–2294 (2018).

40. Zang, X. M. *et al.* A facile method to study the bioaccumulation kinetics of amorphous silica nanoparticles by quantum dot embedding. *Environ Sci Nano* **5**, 2830–2841 (2018).

41. Du, Q. *et al.* Assessment of neurotoxicity induced by different-sized Stöber silica nanoparticles: Induction of pyroptosis in microglia. *Nanoscale* **11**, 12965–12972 (2019).

42. Kamikubo, Y., Yamana, T., Hashimoto, Y. & Sakurai, T. Induction of Oxidative Stress and Cell Death in Neural Cells by Silica Nanoparticles. *ACS Chem Neurosci* **10**, 304–312 (2019).

43. Kim, W. *et al.* A reliable approach for assessing size-dependent effects of silica nanoparticles on cellular internalization behavior and cytotoxic mechanisms. *Int J Nanomedicine* **14**, 7375–7387 (2019).

44. Lee, K. *et al.* Two distinct cellular pathways leading to endothelial cell cytotoxicity by silica nanoparticle size. *J Nanobiotechnology* **17**, 24 (2019).

45. Ren, L. *et al.* Silica nanoparticles induce spermatocyte cell autophagy through microRNA-494 targeting AKT in GC-2spd cells. *Environmental Pollution* **255**, 113172 (2019).

46. Crucho, C. I. C. *et al.* Silica nanoparticles with thermally activated delayed fluorescence for live cell imaging. *Materials Science and Engineering C* **109**, 110528 (2020).

47. Liu, Y. *et al.* Dysfunction of pulmonary epithelial tight junction induced by silicon dioxide nanoparticles via the ROS/ERK pathway and protein degradation. *Chemosphere* **255**, 126954 (2020).

48. Nazarparvar-Noshadi, M., Ezzati Nazhad Dolatabadi, J., Rasoulzadeh, Y., Mohammadian, Y. & Shanehbandi, D. Apoptosis and DNA damage induced by silica nanoparticles and formaldehyde in human lung epithelial cells. *Environmental Science and Pollution Research* **27**, 18592–18601 (2020).

49. Tada-Oikawa, S. *et al.* Functionalized surface-charged sio2 nanoparticles induce pro-inflammatory responses, but are not lethal to caco-2 cells. *Chem Res Toxicol* **33**, 1226–1236 (2020).

50. Wang, M. *et al.* Silica nanoparticles induce lung inflammation in mice via ROS/PARP/TRPM2 signaling-mediated lysosome impairment and autophagy dysfunction. *Part Fibre Toxicol* **17**, 1–22 (2020).

51. Cui, G. *et al.* Activation of Nrf2/HO-1 signaling pathway attenuates ROS-mediated autophagy induced by silica nanoparticles in H9c2 cells. *Environ Toxicol* **36**, 1389–1401 (2021).

52. Hou, S. *et al.* Silica nanoparticles induce mitochondrial pathway-dependent apoptosis by activating unfolded protein response in human neuroblastoma cells. *Environ Toxicol* **36**, 675–685 (2021).

53. Ruan, C. *et al.* An integrative multi-omics approach uncovers the regulatory role of CDK7 and CDK4 in autophagy activation induced by silica nanoparticles. *Autophagy* **17**, 1426–1447 (2021).

54. Hou, S. *et al.* Silica Nanoparticles Cause Activation of NLRP3 Inflammasome in-vitro Model-

Using Microglia. *Int J Nanomedicine* **17**, 5247–5264 (2022).

55. Kim, I. Y. *et al.* Variations in in vitro toxicity of silica nanoparticles according to scaffold type in a 3D culture system using a micropillar/microwell chip platform. *Sens Actuators B Chem* **369**, 132328 (2022).

56. Liang, Q. *et al.* Silica nanoparticles induce hepatocyte ferroptosis and liver injury via ferritinophagy. *Environ Sci Nano* **9**, 3014–3029 (2022).

57. Ma, Y. *et al.* Silica nanoparticles induce pulmonary autophagy dysfunction and epithelial-to-mesenchymal transition via p62/NF-κB signaling pathway. *Ecotoxicol Environ Saf* **232**, 113303 (2022).

58. Zhang, Z. *et al.* Mechanistic study of silica nanoparticles on the size-dependent retinal toxicity in vitro and in vivo. *J Nanobiotechnology* **20**, 1–19 (2022).

59. Hirai, T. *et al.* Cutaneous exposure to agglomerates of silica nanoparticles and allergen results in IgE-biased immune response and increased sensitivity to anaphylaxis in mice. *Part Fibre Toxicol* **12**, 1–6 (2015).

60. Nabeshi, H. *et al.* Systemic distribution, nuclear entry and cytotoxicity of amorphous nanosilica following topical application. *Biomaterials* **32**, 2713–2724 (2011).

61. Lundberg, S. M. & Lee, S. I. A Unified Approach to Interpreting Model Predictions Scott. *Advances in Neural Information Processing Systems 30* **30**, 4768–4777 (2017).

62. Tenzer, S. *et al.* Rapid formation of plasma protein corona critically affects nanoparticle pathophysiology. *Nat Nanotechnol* **8**, 772–781 (2013).

63. Blechinger, J. *et al.* Uptake kinetics and nanotoxicity of silica nanoparticles are cell type dependent. *Small* **9**, 3970–3980 (2013).

64. Lesniak, A. *et al.* Effects of the presence or absence of a protein corona on silica nanoparticle uptake and impact on cells. *ACS Nano* **6**, 5845–5857 (2012).

65. Lesniak, A. *et al.* Nanoparticle adhesion to the cell membrane and its effect on nanoparticle uptake efficiency. *J Am Chem Soc* **135**, 1438–1444 (2013).

66. Francia, V. *et al.* Corona Composition Can Affect the Mechanisms Cells Use to Internalize Nanoparticles. *ACS Nano* **13**, 11107–11121 (2019).

67. Fedeli, C. *et al.* Catastrophic inflammatory death of monocytes and macrophages by overtaking of a critical dose of endocytosed synthetic amorphous silica nanoparticles/serum protein complexes. *Nanomedicine* **8**, 1101–1126 (2013).

68. Pavan, C. *et al.* Nearly free surface silanols are the critical molecular moieties that initiate the toxicity of silica particles. *Proc Natl Acad Sci U S A* **117**, 27836–27846 (2020).

69. Brinker, C. J., Butler, K. S. & Garofalini, S. H. Are nearly free silanols a unifying structural determinant of silica particle toxicity? **117**, 30006-30008 (2020).

70. Zhang, H. *et al.* Processing pathway dependence of amorphous silica nanoparticle toxicity: Colloidal vs pyrolytic. *J Am Chem Soc* **134**, 15790–15804 (2012).

71. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. Catboost: Unbiased boosting with categorical features. *Adv Neural Inf Process Syst* **2018-Decem**, 6638–6648 (2018).

72. Pan, Y. & Kong, L. D. Urate transporter URAT1 inhibitors: a patent review (2012 - 2015). *Expert Opinion on Therapeutic Patents* **26**, 1129-1138 (2016).

73. Dong, Y. *et al.* Novel urate transporter 1 (URAT1) inhibitors: a review of recent patent literature (2016–2019). *Expert Opin Ther Pat* **29**, 871–879 (2019).

74. Shi, X. *et al.* Novel urate transporter 1 (URAT1) inhibitors: a review of recent patent literature

(2020–present). *Expert Opinion on Therapeutic Patents* **32**, 1175–1184 (2022).

75. Zdrazil, B. *et al.* The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res* **52**, D1180-D1192 (2024).
76. Kawashima, H. *et al.* DruMAP: A Novel Drug Metabolism and Pharmacokinetics Analysis Platform. *J Med Chem* **66**, 9697–9709 (2023).
77. Yu, Z., Wing, P. F. & Cheng, C. H. K. Morin (3,5,7,2′,4′-pentahydroxyflavone) exhibits potent inhibitory actions on urate transport by the human urate anion transporter (hURAT1) expressed in human embryonic kidney cells. *Drug Metabolism and Disposition* **35**, 981–986 (2007).
78. Uetake, D. *et al.* Effect of fenofibrate on uric acid metabolism and urate transporter 1. *Internal Medicine* **49**, 89–94 (2010).
79. Wempe, M. F. *et al.* Developing potent human uric acid transporter 1 (hURAT1) inhibitors. *J Med Chem* **54**, 2701–2713 (2011).
80. Shin, H. J. *et al.* Interactions of urate transporter URAT1 in human kidney with uricosuric drugs. *Nephrology* **16**, 156–162 (2011).
81. Wempe, M. F. *et al.* Potent human uric acid transporter 1 inhibitors: In vitro and in vivo metabolism and pharmacokinetic studies. *Drug Des Devel Ther* **6**, 323–339 (2012).
82. Peng, J. *et al.* Discovery of potent and orally bioavailable inhibitors of Human Uric Acid Transporter 1 (hURAT1) and binding mode prediction using homology model. *Bioorg Med Chem Lett* **26**, 277–282 (2016).
83. Pike, A. *et al.* The design, synthesis and evaluation of low molecular weight acidic sulfonamides as URAT1 inhibitors for the treatment of gout. *Medchemcomm* **7**, 1572–1579 (2016).
84. Storer, R. I. *et al.* The discovery and evaluation of diaryl ether heterocyclic sulfonamides as

URAT1 inhibitors for the treatment of gout. *Medchemcomm* **7**, 1587–1595 (2016).

85. Tian, H. *et al.* Discovery of a flexible triazolylbutanoic acid as a highly potent uric acid transporter 1 (URAT1) inhibitor. *Molecules* **21**, 1543 (2016).

86. Cai, W. *et al.* Design, synthesis and biological activity of tetrazole-bearing uric acid transporter 1 inhibitors. *Chem Res Chin Univ* **33**, 49–60 (2017).

87. Yang, X., Pang, X., Fan, L., Li, X. & Chen, Y. Synthesis and evaluation of sulfonamide derivatives as potent Human Uric Acid Transporter 1 (hURAT1) inhibitors. *Bioorg Med Chem Lett* **27**, 1919–1922 (2017).

88. Zhang, X. *et al.* Discovery of Flexible Naphthyltriazolylmethane-based Thioacetic Acids as Highly Active Uric Acid Transporter 1 (URAT1) Inhibitors for the Treatment of Hyperuricemia of Gout. *Med Chem (Los Angeles)* **13**, 260–281 (2016).

89. Cai, W. *et al.* Systematic structure-activity relationship (SAR) exploration of diarylmethane backbone and discovery of a highly potent novel uric acid transporter 1 (URAT1) inhibitor. *Molecules* **23**, 252 (2018).

90. Tashiro, Y. *et al.* Effects of osthol isolated from Cnidium monnieri fruit on urate transporter 1. *Molecules* **23**, 2837 (2018).

91. Li, X., Liu, J., Ma, L. & Fu, P. Pharmacological urate-lowering approaches in chronic kidney disease. *European Journal of Medicinal Chemistry* **166**, 186–196 (2019).

92. Wu, J. wei *et al.* Synthesis, biological evaluation and 3D-QSAR studies of 1,2,4-triazole-5-substituted carboxylic acid bioisosteres as uric acid transporter 1 (URAT1) inhibitors for the treatment of hyperuricemia associated with gout. *Bioorg Med Chem Lett* **29**, 383–388 (2019).

93. Saito, H. *et al.* Omega-3 polyunsaturated fatty acids inhibit the function of human URAT1, a

renal urate re-absorber. *Nutrients* **12**, 1601 (2020).

94. Toyoda, Y. *et al.* Inhibitory effect of Citrus flavonoids on the in vitro transport activity of human urate transporter 1 (URAT1/SLC22A12), a renal re-absorber of urate. *NPJ Sci Food* **4**, 3 (2020).

95. Uda, J. *et al.* Discovery of Dotinurad (FYU-981), a New Phenol Derivative with Highly Potent Uric Acid Lowering Activity. *ACS Med Chem Lett* **11**, 2017–2023 (2020).

96. Zhao, T. *et al.* Novel Human Urate Transporter 1 Inhibitors as Hypouricemic Drug Candidates with Favorable Druggability. *J Med Chem* **63**, 10829–10854 (2020).

97. Chen, X. *et al.* Novel natural scaffold as hURAT1 inhibitor identified by 3D-shape-based, docking-based virtual screening approach and biological evaluation. *Bioorg Chem* **117**, 105444 (2021).

98. Zhou, H. *et al.* Development of a fluorescence-based assay for screening of urate transporter 1 inhibitors using 6-carboxyfluorescein. *Anal Biochem* **626**, 114246 (2021).

99. Zhang, J. *et al.* Design, synthesis and activity evaluation of novel lesinurad analogues containing thienopyrimidinone or pyridine substructure as human urate transporter 1 inhibitors. *Eur J Med Chem* **244**, 114816 (2022).

100. Zhao, Z. *et al.* Discovery of novel verinurad analogs as dual inhibitors of URAT1 and GLUT9 with improved Druggability for the treatment of hyperuricemia. *Eur J Med Chem* **229**, 114092 (2022).

101. Zhao, Z. *et al.* Discovery of novel benzbromarone analogs with improved pharmacokinetics and benign toxicity profiles as antihyperuricemic agents. *Eur J Med Chem* **242**, 114682 (2022).

102. Bento, A. P. *et al.* An open source chemical structure curation pipeline using RDKit. *J Cheminform* **12**, 51 (2020).

103. Moriwaki, H., Tian, Y. S., Kawashita, N. & Takagi, T. Mordred: A molecular descriptor calculator. *J Cheminform* **10**, 4 (2018).

104. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* **2**, 56–67 (2020).

105. Cawley, G. C. & Talbot, N. L. C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research* **11**, 2079–2107 (2010).

106. Wellawatte, G. P., Seshadri, A. & White, A. D. Model agnostic generation of counterfactual explanations for molecules. *Chem Sci* **13**, 3697–3705 (2022).

107. Vogt, M., Stumpfe, D., Maggiora, G. M. & Bajorath, J. Lessons learned from the design of chemical space networks and opportunities for new applications. *J Comput Aided Mol Des* **30**, 191–208 (2016).

108. Callis, T. B., Garrett, T. R., Montgomery, A. P., Danon, J. J. & Kassiou, M. Recent Scaffold Hopping Applications in Central Nervous System Drug Discovery. *Journal of Medicinal Chemistry* **65,** 13483–13504 (2022).

109. Banerjee, P., Eckert, A. O., Schrey, A. K. & Preissner, R. ProTox-II: A webserver for the prediction of toxicity of chemicals. *Nucleic Acids Res* **46**, W257–W263 (2018).