A web-based multi-target cytotoxicity prediction for multi-component nanoparticles: nano-QSAR model with extended applicability domain

Jaehyeon Park[a, b], Hyun Kil Shin[a, b, *]

[a] Department of Predictive Model Research, Korea Institute of Toxicology, Daejeon 34114, Republic of Korea
[b] Human and Environmental Toxicology, University of Science and Technology, Daejeon, 34114, Republic of Korea

Abstract

As nanotechnology advances, increasingly complex nanoparticles are being developed for various applications, raising critical concerns about their potential toxicity. Not only Nano-QSAR models have been developed to predict their toxicity by cell lines separately, but also their applicability domain (AD) has been limited to specific nanoparticle types (i.e., bare metal oxide, coated metal, or carbon-based nanomaterials). This research introduced multi-target nano-QSAR model, being developed with improved AD by training the model on multi-component nanoparticles (MC NPs) to use size-dependent electron configuration fingerprint (SDEC FP) and with one-hot encoded cell features to predict cytotoxicity of MC NPs over 110 cell lines. The CatBoost regression model showed good performance (R^2 test = 0.877) and is now accessible through user friendly web interface (https://www.kitox.re.kr/nanotoxradar). NanoToxRadar allows users to input nanoparticle specifications-including core, shell, doping, and coating materials, along with particle diameter-and receive predicted pIC_{50} values across 110 cell lines.

Introduction

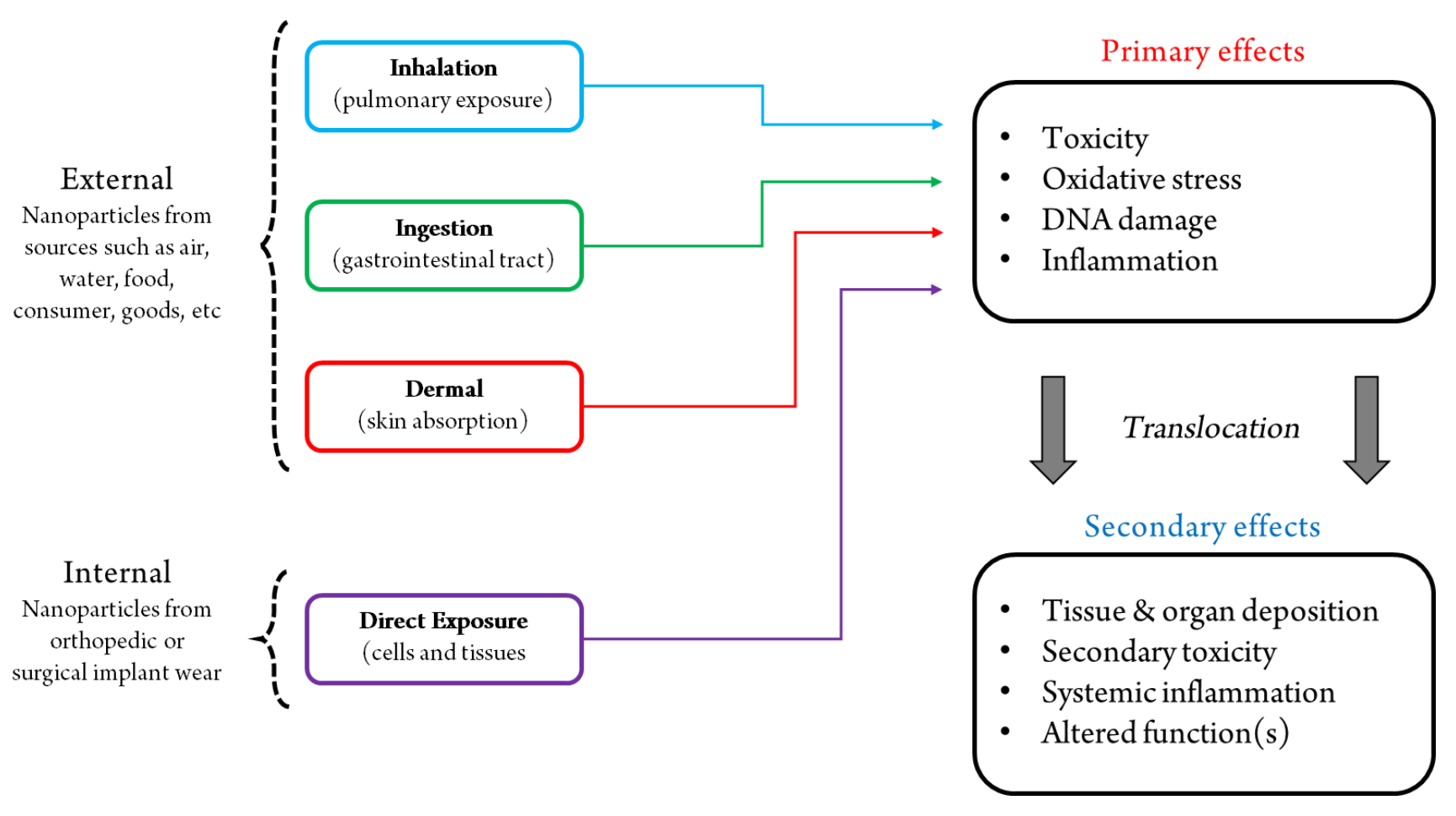


Figure 1) The block diagram shows effects steps and the potential hazard of nanoparticles with external and internal way.

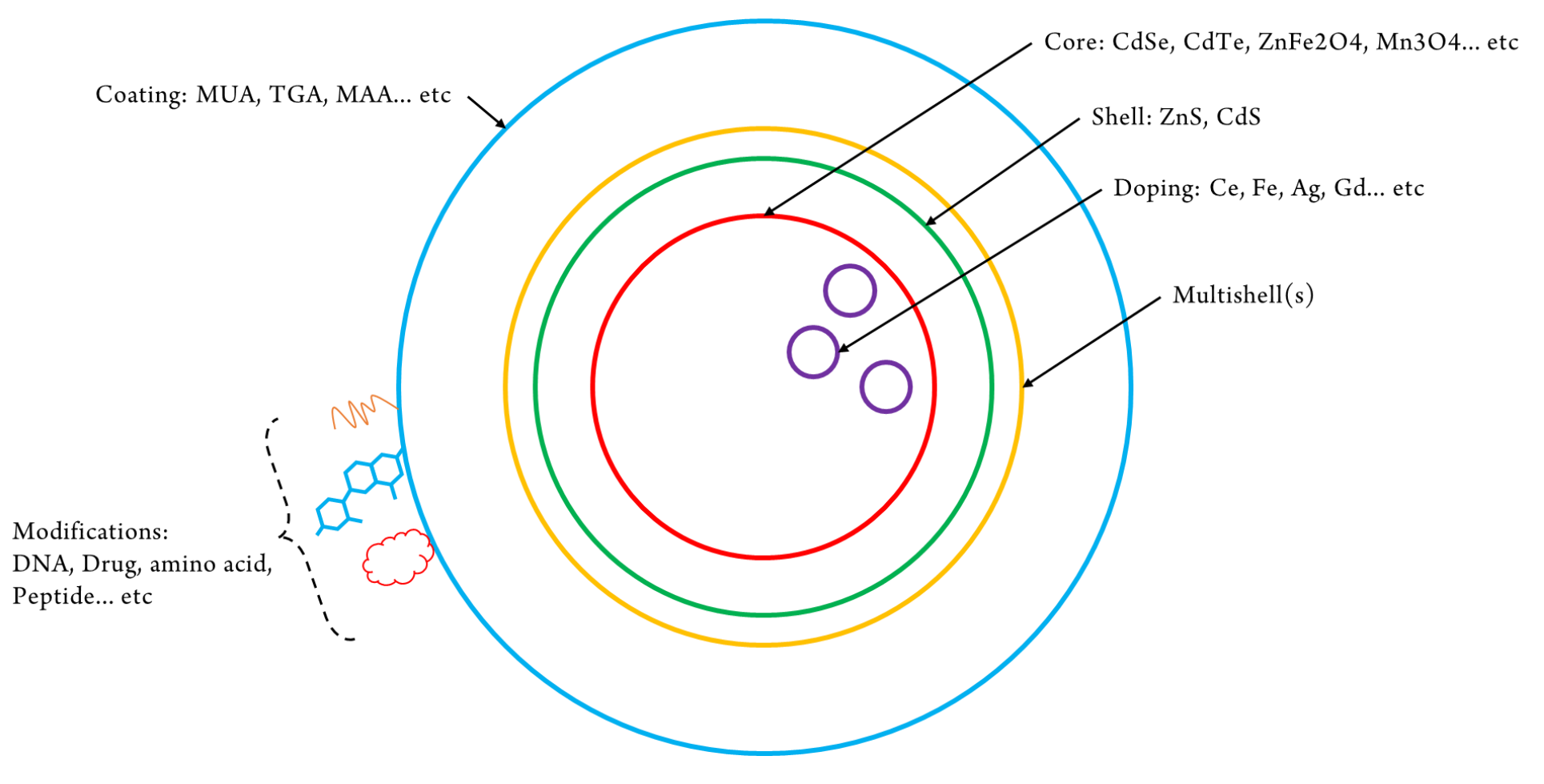

Coating: MUA, TGA, MAA... etc
Core: CdSe, CdTe, ZnFe2O4, Mn3O4... etc
Shell: ZnS, CdS
Doping: Ce, Fe, Ag, Gd... etc
Multishell(s)
Modifications:
DNA, Drug, amino acid,
Peptide... etc

Figure 2) Schematic of multi-component nanoparticles (MC NPs) including core, shell and coating with modification.

Significance of the Study

- The interest is shifted towards the potential toxicity of multi-component nanoparticles (MC NPs, Figure 2) due to their small size, large surface area per volume, and even their complex components as nanotechnology advances.

Main Problem

1. Existing Nano-QSAR models have a restricted applicability domain (AD) due to the scarcity of comprehensive nanotoxicity data available for model development.
2. While quantum mechanical (QM) and molecular dynamics (MD) descriptors offer theoretical advantages, they require substantial computational resources, additionally, molecular clusters representing nanomaterials often suffer from poor reproducibility.
3. Many nano-QSAR models have been developed separately, targeting specific endpoint such as cytotoxicity in specific cell lines.

Suggested Solution

1. Application of size-dependent electron configuration fingerprint (SDEC FP) to represent MC NP structures, improving model's AD.
2. Multi-target prediction is a better approach to increase data size through integration of different target endpoints measured from 110 cell types by introducing cell features.
3. The optimal model is deployed on web environment to easily access of the model to the research community.

Methods

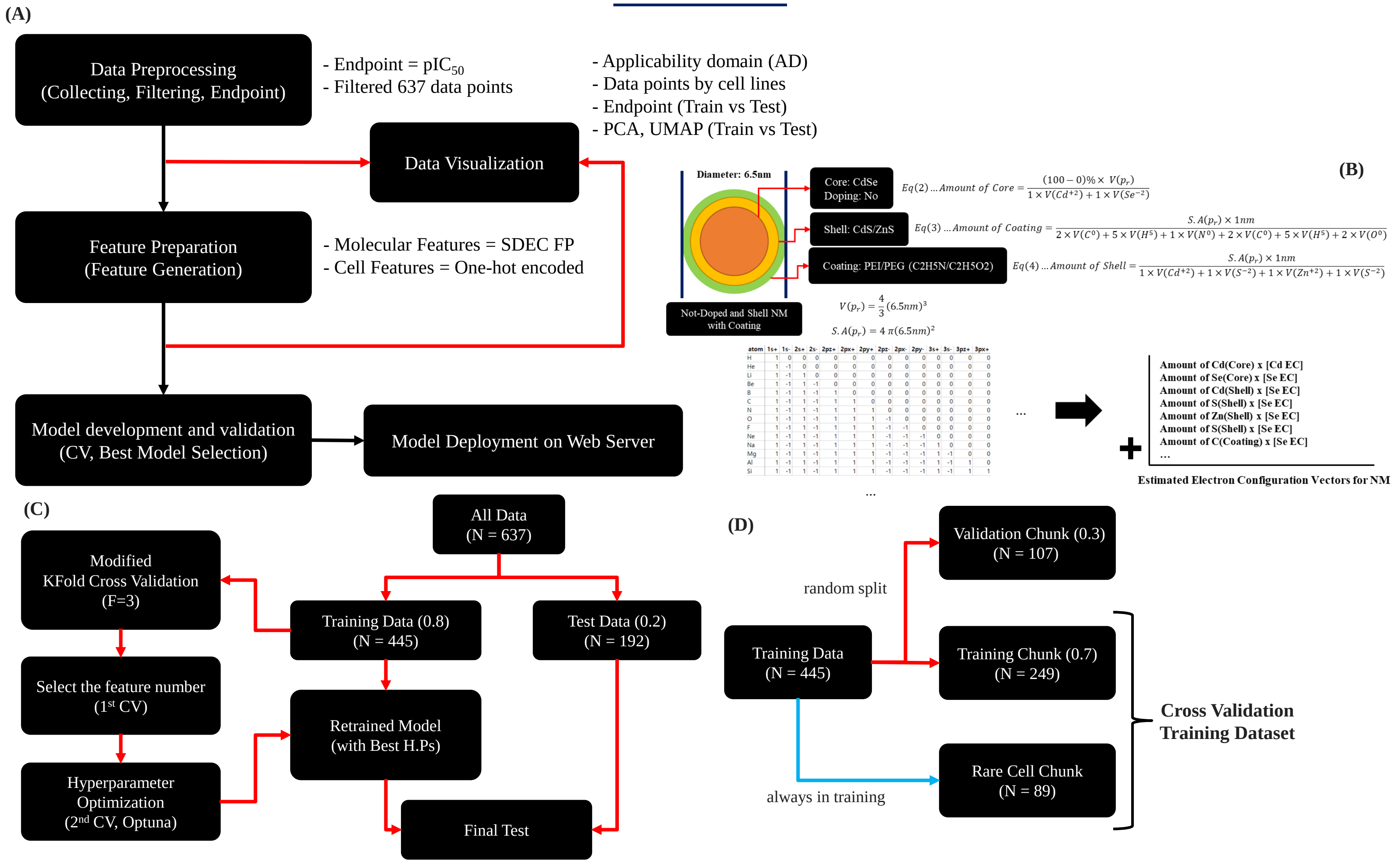

(A)
Data Preprocessing
(Collecting, Filtering, Endpoint)
- Endpoint = pIC_{50}
- Filtered 637 data points
Data Visualization
- Applicability domain (AD)
- Data points by cell lines
- Endpoint (Train vs Test)
- PCA, UMAP (Train vs Test)
Feature Preparation
(Feature Generation)
- Molecular Features = SDEC FP
- Cell Features = One-hot encoded
Model development and validation
(CV, Best Model Selection)
Model Deployment on Web Server
(B)
Diameter: 6.5nm
Core: CdSe
Doping: No
Shell: CdS/ZnS
Coating: PEI/PEG (C2H5N/C2H5O2)
Not-Doped and Shell NM with Coating
$Eq(2) \ldots Amount\ of\ Core = \frac{(100-0)\% \times V(p_r)}{1 \times V(Cd^{+2}) + 1 \times V(Se^{-2})}$
$Eq(3) \ldots Amount\ of\ Coating = \frac{S.A(p_r) \times 1nm}{2 \times V(C^0) + 5 \times V(H^5) + 1 \times V(N^0) + 2 \times V(C^0) + 5 \times V(H^5) + 2 \times V(O^0)}$
$Eq(4) \ldots Amount\ of\ Shell = \frac{S.A(p_r) \times 1nm}{1 \times V(Cd^{+2}) + 1 \times V(S^{-2}) + 1 \times V(Zn^{+2}) + 1 \times V(S^{-2})}$
$V(p_r) = \frac{4}{3}(6.5nm)^3$
$S.A(p_r) = 4\pi(6.5nm)^2$
Amount of Cd(Core) x [Cd EC]
Amount of Se(Core) x [Se EC]
Amount of Cd(Shell) x [Se EC]
Amount of S(Shell) x [Se EC]
Amount of Zn(Shell) x [Se EC]
Amount of S(Shell) x [Se EC]
Amount of C(Coating) x [Se EC]
Estimated Electron Configuration Vectors for NM
(C)
All Data
(N = 637)
Training Data (0.8)
(N = 445)
Test Data (0.2)
(N = 192)
Modified
KFold Cross Validation
(F=3)
Select the feature number
(1st CV)
Hyperparameter
Optimization
(2nd CV, Optuna)
Retrained Model
(with Best H.Ps)
Final Test
(D)
Training Data
(N = 445)
random split
always in training
Validation Chunk (0.3)
(N = 107)
Training Chunk (0.7)
(N = 249)
Rare Cell Chunk
(N = 89)
Cross Validation
Training Dataset

Figure 3) Pipeline diagram of entire model development. (A) Model development pipeline diagram from data preprocessing to model deployment. (B) SDEC FP calculation schematic diagram with simple example in dataset. (C) Detailed pipeline diagram of model development and validation from data separation to final test. (D) Detailed pipeline diagram of modified KFold cross validation in selection of feature number and hyperparameter optimization (K=3, folding three times in above way)

- Calculation of SDEC FP for the MC NPs as follows:
  1) full size of SDEC FP without compression
  2) aggregated SDEC FP by adding up atomic orbital indices in the identical energy level theoretically
  3) the aggregated SDEC FP without positive and negative sign, ignoring spin number
- One-hot encoded cell information vectors as follows:
  1) all five-cell information
  2) cell name alone
  3) cell name and source tissues/organs
  4) cell name and anatomical classification

Results

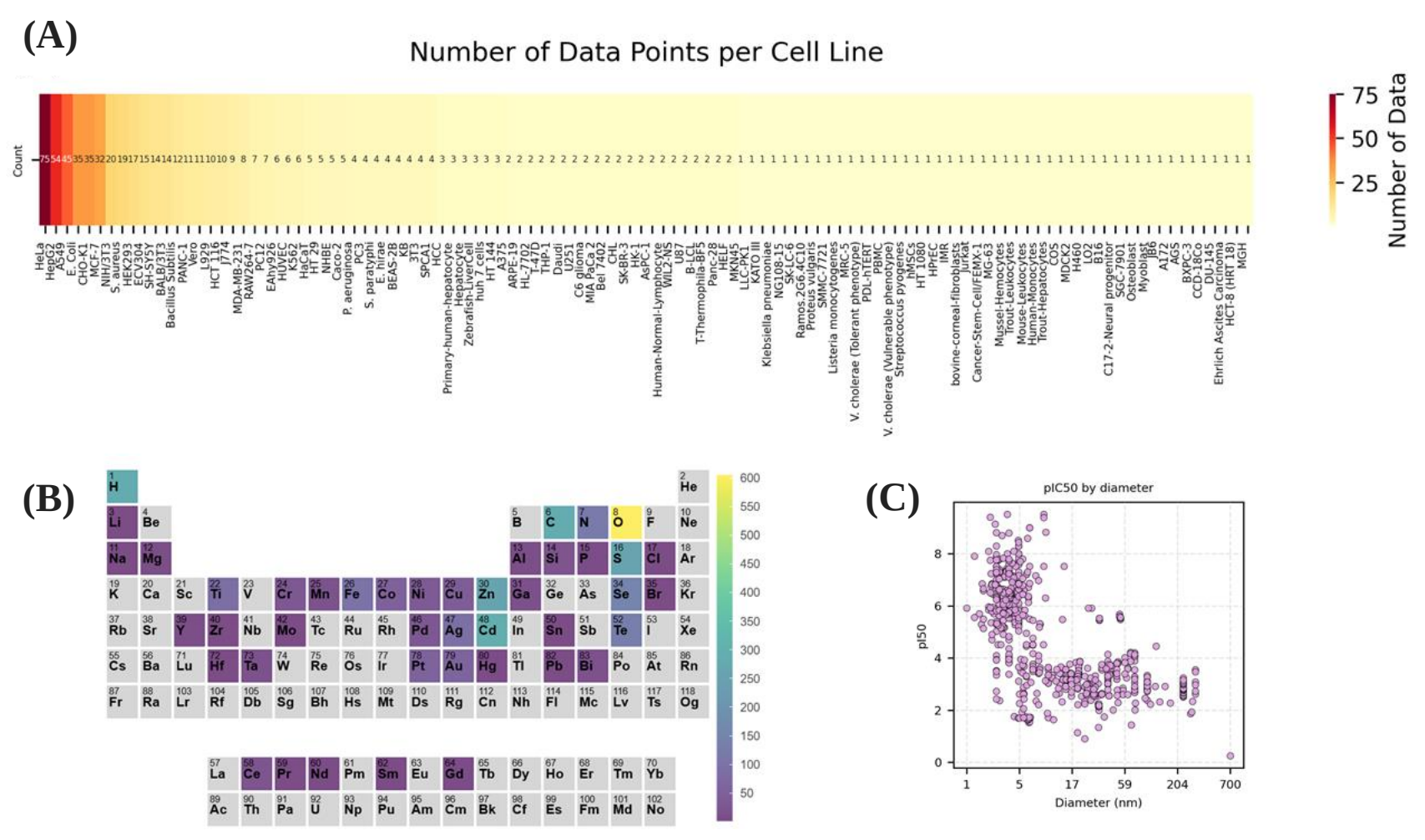


Figure 4) Heatmap visualizes the distribution of different cell lines across data points, with color intensity indicating the frequency of cells. (A). Periodic table highlights compositional complexity of MC-NPs in the dataset with color gradient indicating the frequency of each element's occurrence (B). Scatter plot demonstrates the relationship between diameter of MC-NPs (1-700 nm) and cytotoxicity (pIC50). Higher cytotoxicity was observed among MC-NPs with diameter smaller than 5nm. (C)

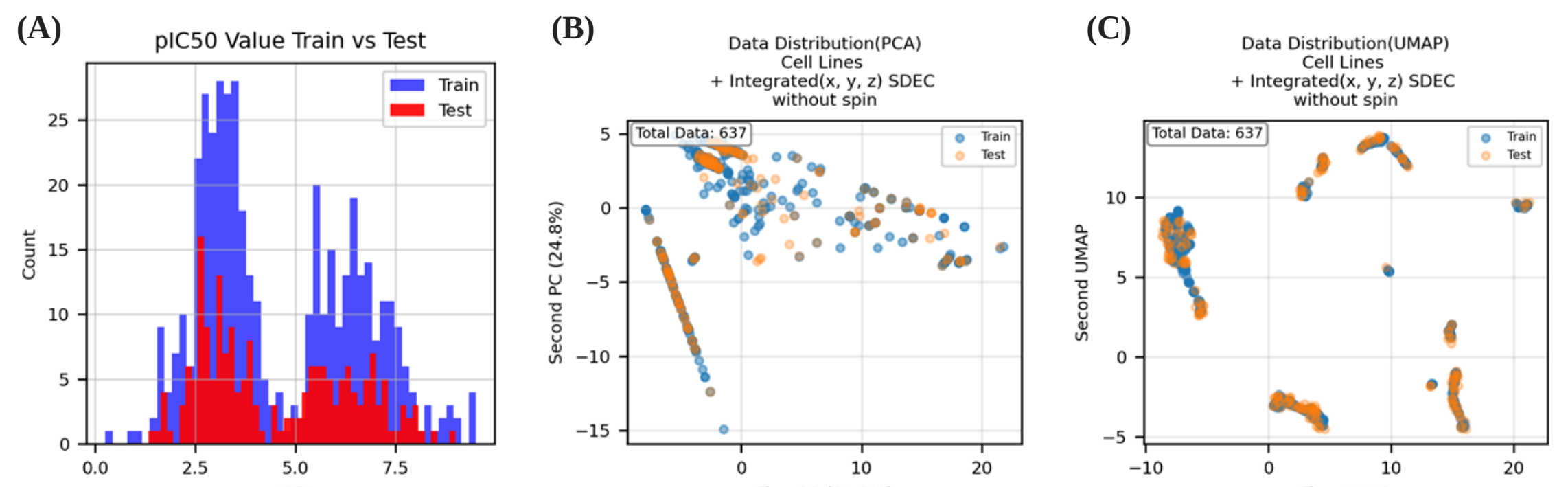


Figure 5) Data distribution for the endpoint and feature space visualization: training vs test dataset. Histogram showing the distribution of endpoints (pIC50) across train (blue) and test (red) datasets, demonstrating balanced representation of endpoints. (B) Principal Component Analysis (PCA) visualization of the 130-feature dataset. Train (blue) and test (orange) sets show similar distribution patterns. (C) Uniform Manifold Approximation and Projection (UMAP) plot for the 130-feature dataset revealing the underlying structure of the data in two dimensions, with consistent distribution patterns between train (blue) and test (orange) datasets.

Data Exploration

- 110 cells are included in the dataset, among which 68 cells were tested for only one or two MC NPs, thereby accounting for 89 data points were used in the training set only. (Figure 4A)
- The most common elements of MC NPs in the dataset are oxygen, hydrogen, carbon, and sulfur since many MC-NPs are coated NPs, also Zn and Cd are common elements in the dataset since they are commonly in coatings, shells, dopants, and core materials, which showed AD for the developed model. (Figure 4B)
- The endpoint, which is pIC_{50} , increases exponentially with decreasing MC NP size. (Figure 4C) Also, the distribution of endpoint between training and test data revealed a similar pattern. (Figure 5A)
- The smallest feature was 130 in size (aggregated SDEC FP without spin and cell name one-hot encoded vectors), which showed good feature space similarity between the training and test dataset, visualized using PCA and UMAP. (Figure 5B, 5C)

Table 1) Performance metrics of machine learning models with different feature combinations.

Number of features	Model	$RMSE_{CV}$	R^2_{CV}	$RMSE_{Test}$	R^2_{Test}	$RMSE_{Test}$ over endpoint range (%)	Feature description*
314	CatBoost	0.602 ± 0.013	0.903 ± 0.005	0.623	0.886	6.49%	All cell information & SDEC FP
	ExtraTrees	0.818 ± 0.032	0.820 ± 0.019	0.704	0.855	8.83%	
	SVR	0.691 ± 0.038	0.871 ± 0.021	0.633	0.883	7.46%	
	XGBoost	0.638 ± 0.073	0.889 ± 0.026	0.672	0.868	6.88%	
	GBR	0.662 ± 0.022	0.883 ± 0.007	0.776	0.823	7.14%	
	RandomForest	0.776 ± 0.021	0.839 ± 0.012	0.705	0.854	8.37%	
	MLP	0.743 ± 0.009	0.852 ± 0.011	0.667	0.87	8.02%	
	Transformer	0.717 ± 0.046	0.861 ± 0.026	0.72	0.848	7.74%	
150	CatBoost	0.652 ± 0.047	0.885 ± 0.017	0.703	0.855	7.04%	Cell names & aggregated SDEC FP
	ExtraTrees	0.885 ± 0.020	0.790 ± 0.011	0.75	0.835	9.55%	
	SVR	**0.727 ± 0.033**	**0.857 ± 0.020**	**0.617**	**0.888**	**7.85%**	
	XGBoost	0.698 ± 0.059	0.869 ± 0.021	0.691	0.86	7.53%	
	GBR	0.698 ± 0.072	0.868 ± 0.028	0.69	0.86	7.53%	
	RandomForest	0.833 ± 0.015	0.814 ± 0.002	0.736	0.841	8.99%	
	MLP	0.778 ± 0.048	0.837 ± 0.025	0.76	0.831	8.39%	
	Transformer	0.806 ± 0.049	0.824 ± 0.030	0.673	0.867	8.69%	
130	**CatBoost**	**0.691 ± 0.029**	**0.872 ± 0.005**	**0.649**	**0.877**	**7.45%**	Cell names & aggregated SDEC FP without spin
	ExtraTrees	0.987 ± 0.039	0.739 ± 0.013	0.817	0.804	10.65%	
	SVR	**0.775 ± 0.013**	**0.839 ± 0.012**	**0.635**	**0.882**	**8.36%**	
	XGBoost	0.725 ± 0.024	0.859 ± 0.006	0.697	0.858	7.83%	
	GBR	0.706 ± 0.040	0.866 ± 0.010	0.703	0.855	7.62%	
	RandomForest	0.859 ± 0.008	0.802 ± 0.011	0.728	0.845	9.27%	
	MLP	0.822 ± 0.044	0.817 ± 0.029	0.744	0.838	8.87%	
	Transformer	0.816 ± 0.022	0.821 ± 0.018	0.74	0.84	8.80%	

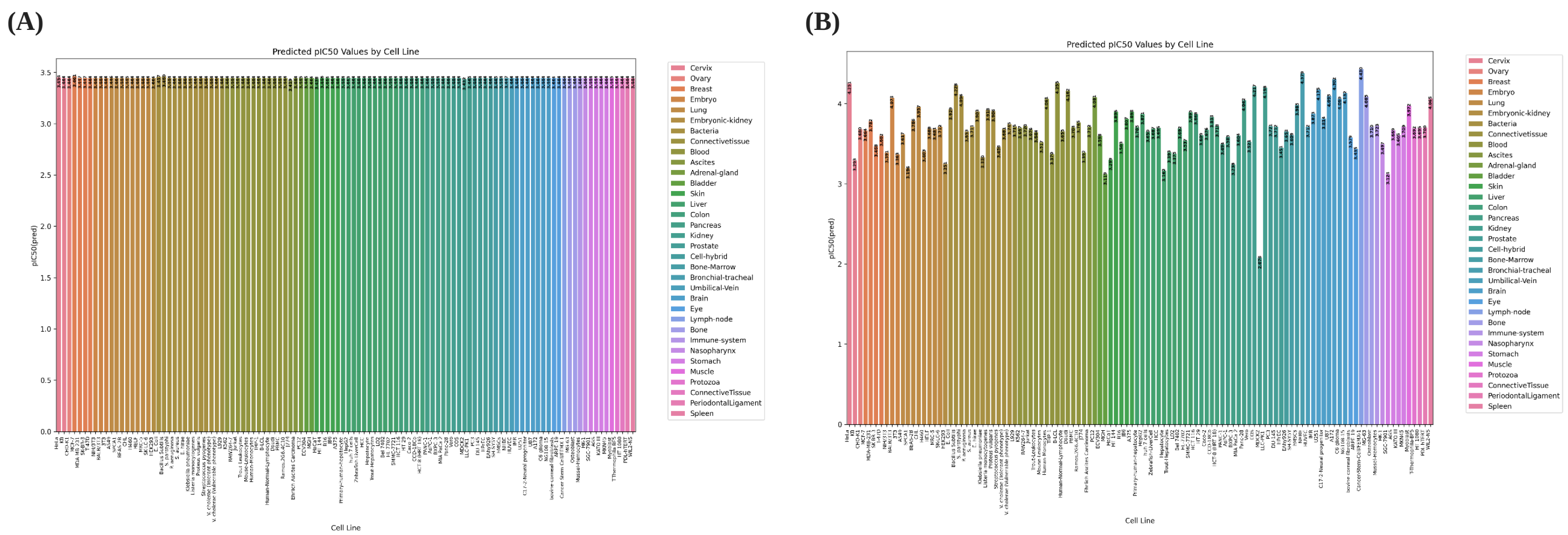


Figure 6) Predicted pIC_{50} values by cell lines (A) Prediction value of SVR with 150 features according to 110 cell lines. (B) Prediction value of CatBoost with 130 features according to 110 cell lines.

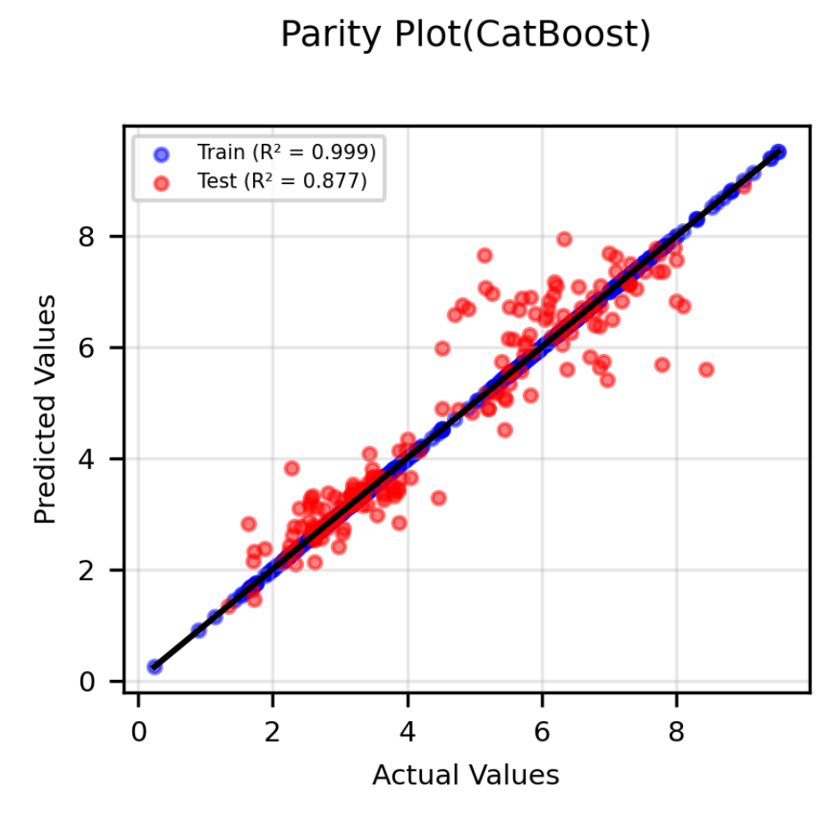
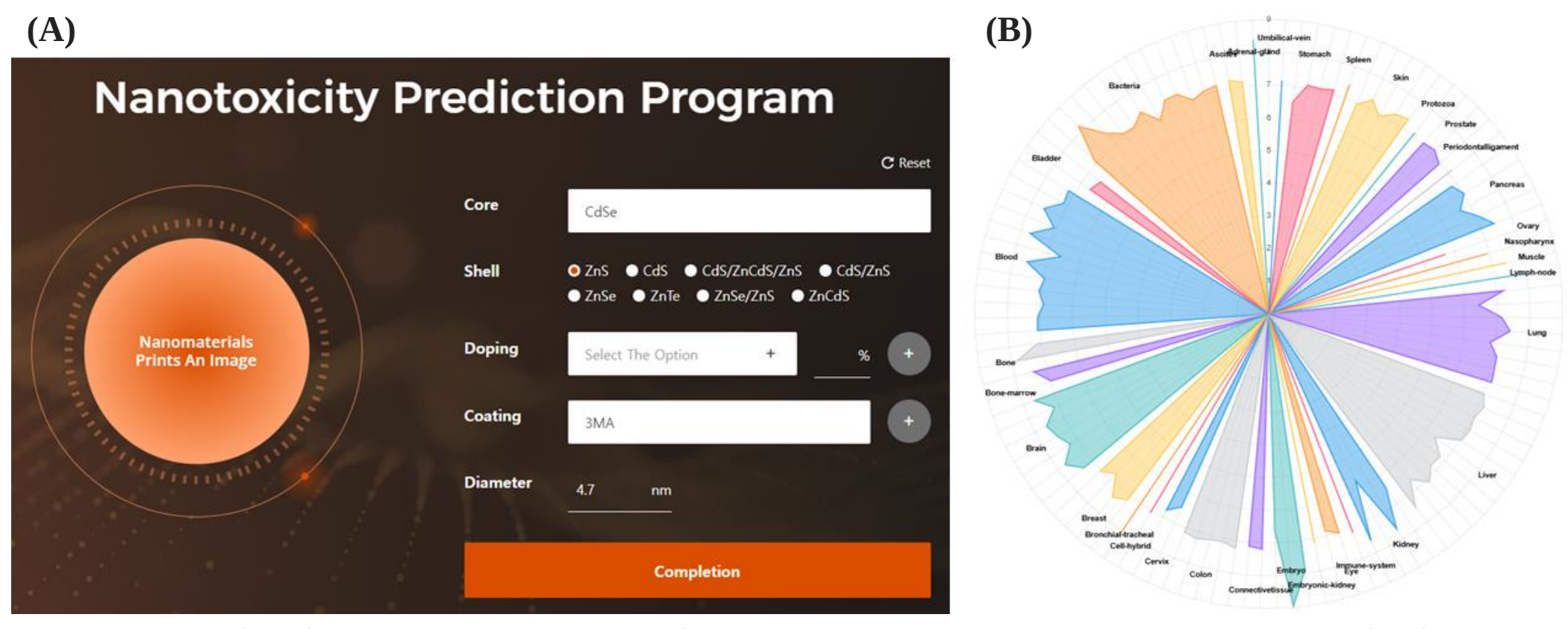


Figure 7) Parity plot comparing predicted and experimental values for the CatBoost regressor model with 130 features.

Model Development

- According to Table 1, SVR with 150 features achieved the best prediction accuracy (R^2_{Test} = 0.888) even with 130 features (R^2_{Test} = 0.882), avoiding overfitting.
- However the SVR models yielded very similar predicted values across different cell types for most MC NPs (Figure 6A), which indicates that SVR failed to learn discrepancies among cell information.
- The CatBoost model with 130 features not only delivered the second-highest performance (R^2_{Test} = 0.877, Table 1, Figure 7) after SVR but also the CatBoost model was also capable of predicting differences in pIC_{50} values across different cell types (Figure 6B), which it was selected as the optimal model and deployed as the web service.



Nanotoxicity Prediction Program

Core: CdSe

Shell: ● ZnS ● CdS ● CdS/ZnCdS/ZnS ● CdS/ZnS ● ZnSe ● ZnTe ● ZnSe/ZnS ● ZnCdS

Doping: Select The Option %

Coating: 3MA

Diameter: 4.7 nm

Figure 8) Web interface of NanoToxRadar and distribution of nanotoxicity prediction results across cell types. (A) User interface for query NP include core, shell, doping, and coating compositions with doping ratio and diameter. (B) Radar plot shows the distribution of pIC50.

Model Deployment

- SDEC FP doesn't require a high computational cost while major obstacles for model deployment are the high computational cost for QM or DM descriptor preparation.
- The simplicity of SDEC FP produces identical descriptor values for the identical MC NPs, which means that the predicted values of the model are highly reproducible.
- NanoToxRadar is developed under responsive web design, thus researchers can use the model on the mobile environment as well. (Figure 8)

Acknowledgements

This work was financially supported by the Ministry of Trade, Industry and Energy (MOTIE) and Korea Institute for Advancement of Technology (KIAT) through the International Cooperative R&D program (P0019147), as well as the Research Program for Agriculture Science and Technology Development (Project No. 00400007) from the National Institute of Agricultural Sciences, Rural Development Administration. The research was conducted in cooperation with SUNSHINE (funded by the European Union's HORIZON 2020 program under grant agreement no. 952924) and the Explainable AI for Molecules - AiChemist project (funded by the Horizon Europe Marie Skłodowska-Curie Actions Doctoral Network under grant agreement number 101120466).