Introduction of NanoToxRadar

Drug Consumption by Gut-Microbiome

DC14 – Jaehyeon Park[1,2]

[1]Human and Environmental Toxicology, UST (KIT school)
[2]Predictive Model Research Center, Korea Institute of Toxicology

KIT 국가독성과학연구소 Korea Institute of Toxicology

Introduction

Simple introduction about me (Background field, hobbies, ...etc.)

Korean Name: Jaehyeon Park

English Name: Lukas Park

2024 **AiChemist**
DC 14 Fellows

2024 **UST**
Major: Human and Environmental Toxicology

Surgery for Slipped disc

2019 **CNU**
Major: Physics
Minor: Chemical Engineering & Applied Chemistry

2014 **DFLHS**
Major: German

Hobbies

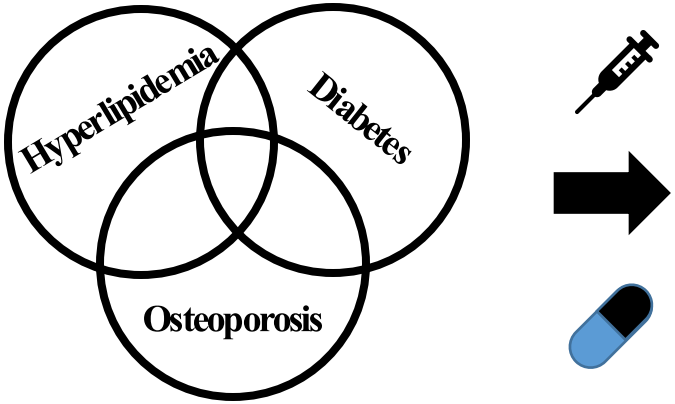
Workout

Movies

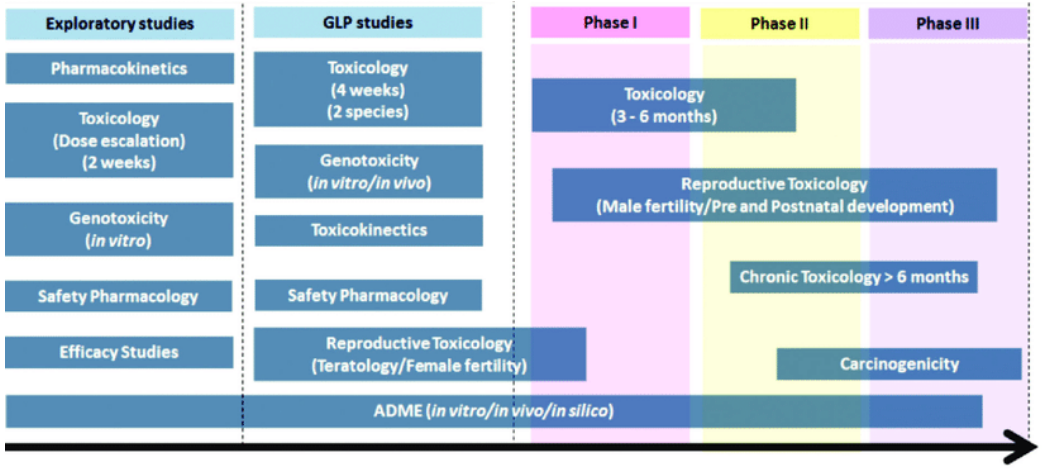
Introduction

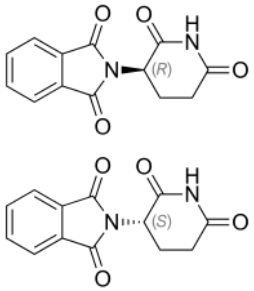
Main Research Field

Limitation



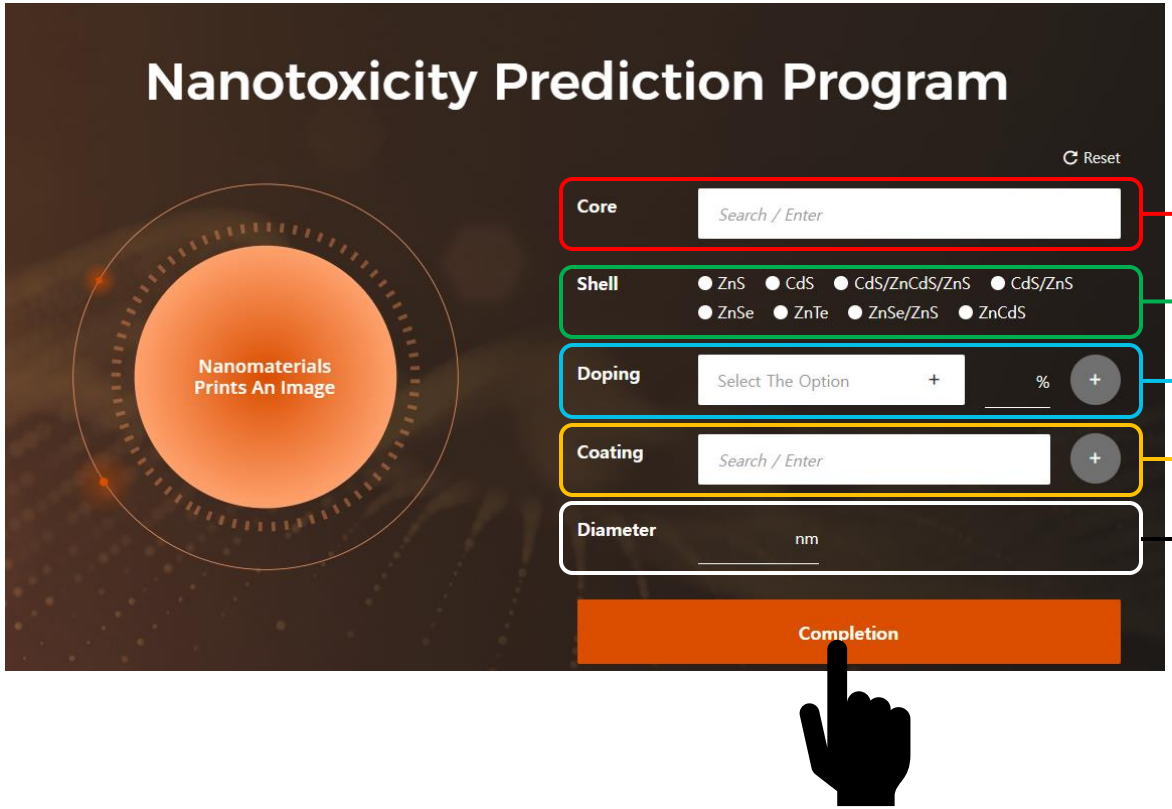
Toxicology in Drug Development





NanoToxRadar

NanoToxRadar



Select or Type the "Core"

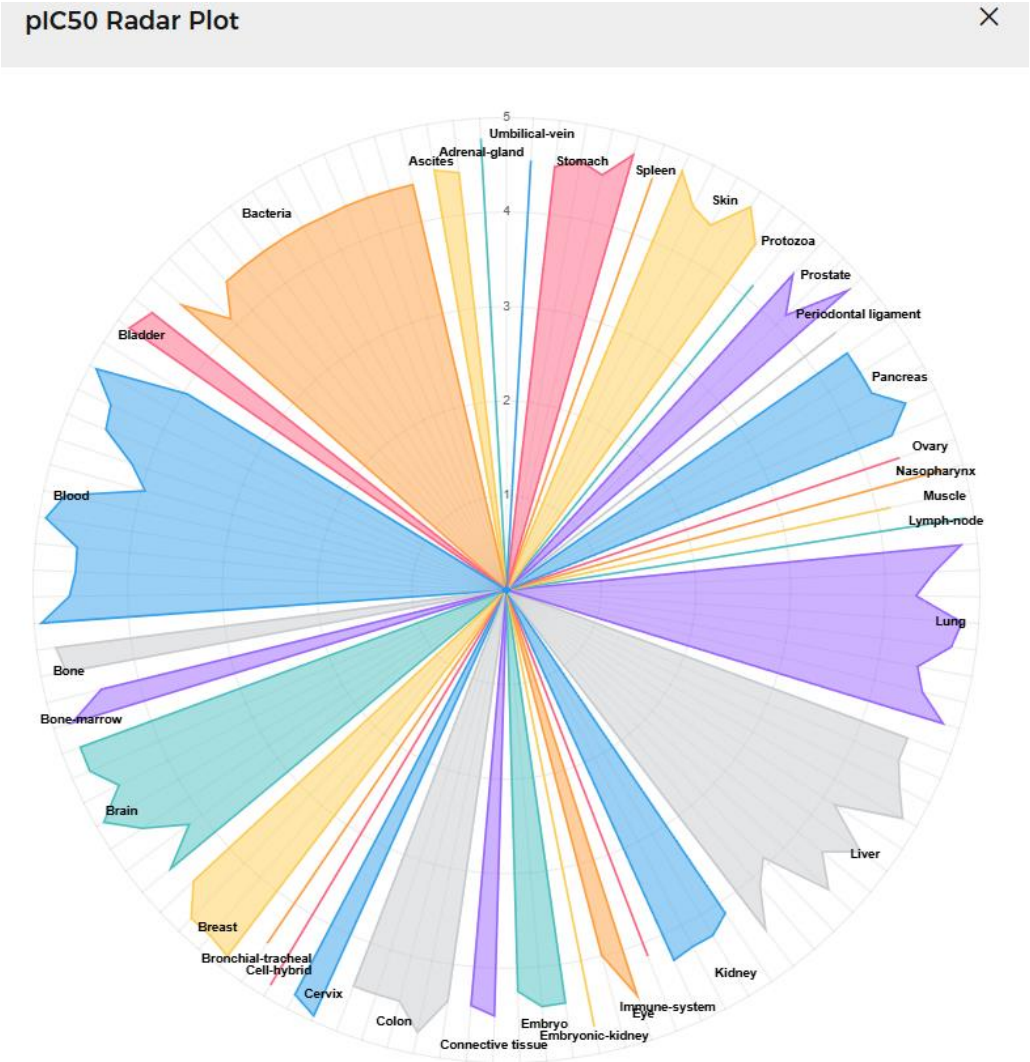
Select the "Shell"

Select the "Doping" with typing Ratio

Select or Type the "Coating"

Type the "Diameter"

NanoToxRadar



Background

[Background]

Nanotoxicity

Hazard of nanoparticles
(even potential harmful effects)

1. Small size
2. Large surface area per volume

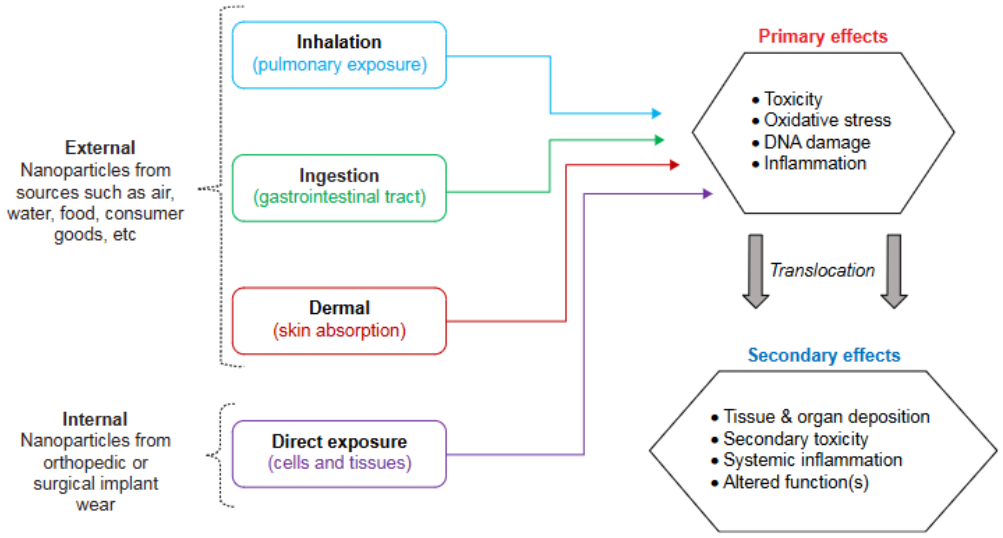


Figure 1 Routes and potential detrimental effects of nanoparticle exposure.

Fig Ref) Armstead, A. L., & Li, B. (2016). Nanotoxicity: emerging concerns regarding nanomaterial safety and occupational hard metal (WC-Co) nanoparticle exposure. International Journal of Nanomedicine, 11, 6421–6433. https://doi.org/10.2147/IJN.S121

Background

[Background]

Multi-Components NanoParticls (MC-NP)

MC-NP

1. Modification of surface area (coating)
2. Multi Doped
3. Add a shell components to NPs
4. Limited applicability domain (AD) for nanotoxicity prediction

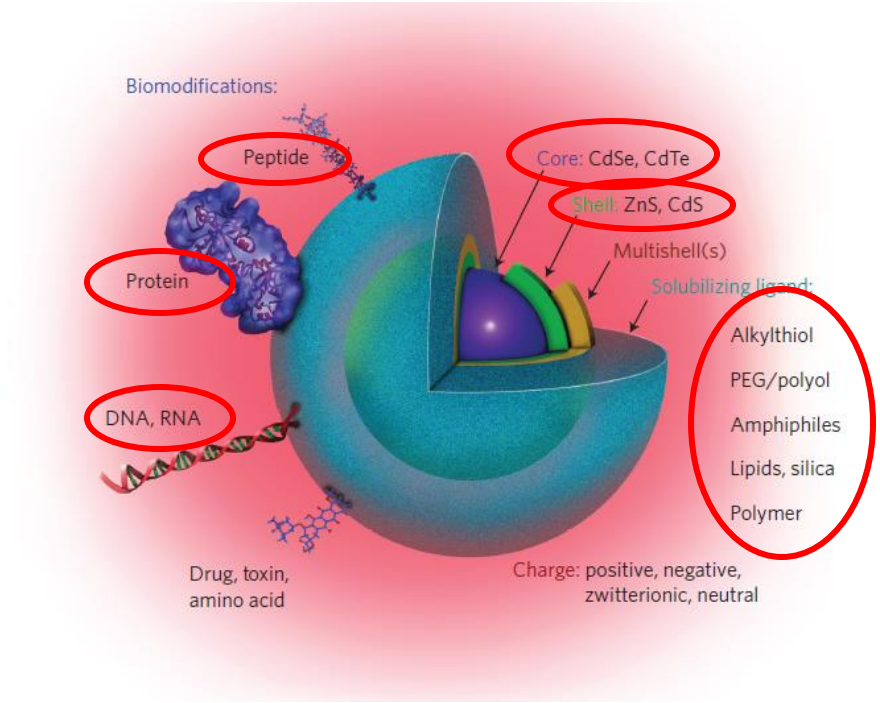
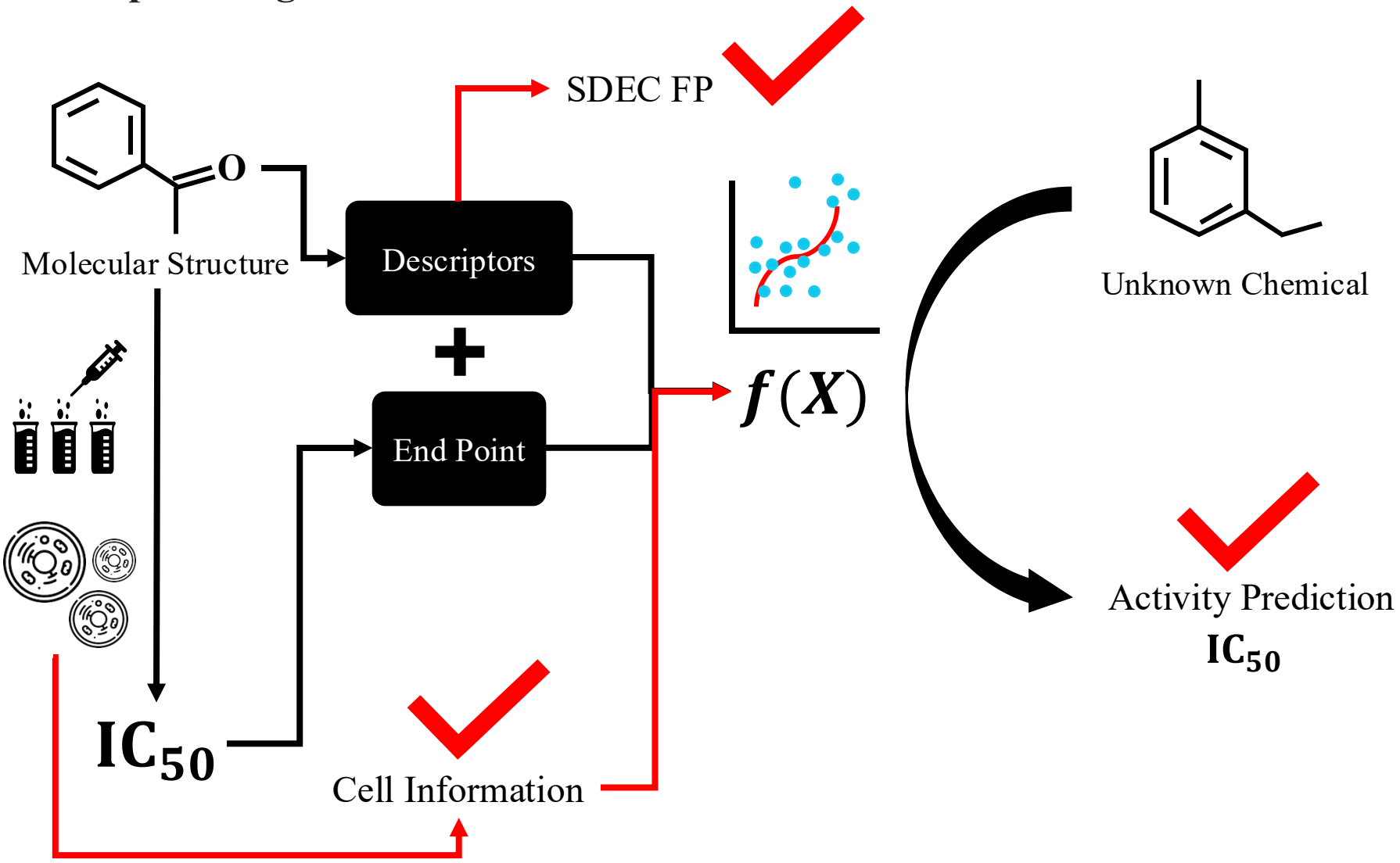


Fig Ref) Oh, E., Liu, R., Nel, A. *et al.* Meta-analysis of cellular toxicity for cadmium-containing quantum dots. *Nature Nanotech* **11**, 479–486 (2016). https://doi.org/10.1038/nnano.2015.338

Data Preprocessing



Model Development

All Data

Training Data (0.8)

Test Data (0.2)

Modified KFold Cross-Validation (F=3)

Select the number of features (1st CV)

Hyperparameter Optimization (2nd CV)

Retrained Model (with Best H.Ps)

Final Test

Condition
the specific training data must not be included in validation and test but in training.

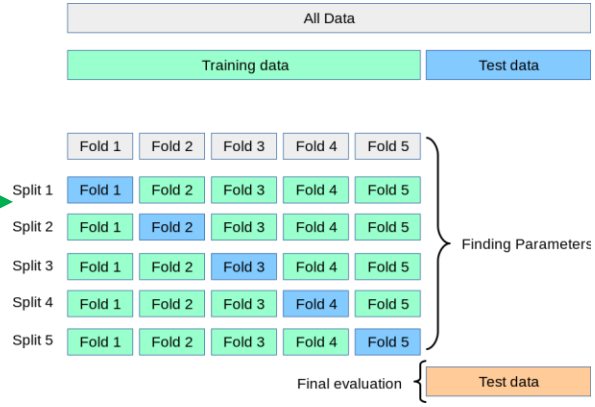


Fig Ref) https://scikit-learn.org/1.5/modules/cross_validation.html

Data Preprocessing

New Data (Sunshine + Q.D) 921 samples

Nano-Particles

- Core
- Doping (%)
- Shell
- Coating

New Composition

More Complicated Core Compositions

Previous	Now
ZnO TiO2 …	Zn0.11Cu0.89O Co0.85Bi0.15Fe2O4 …

Two or More Doping and Coating and More Complicated Coating Compositions

Previous	Now
Co (1~5%) Fe(1~5%) …	Ag(7.4%)/Cu(2.1%)/Co(4%) Fe3O4(50%) C16H35N/CH3(CH2)nCOOH …

Nano-Particles

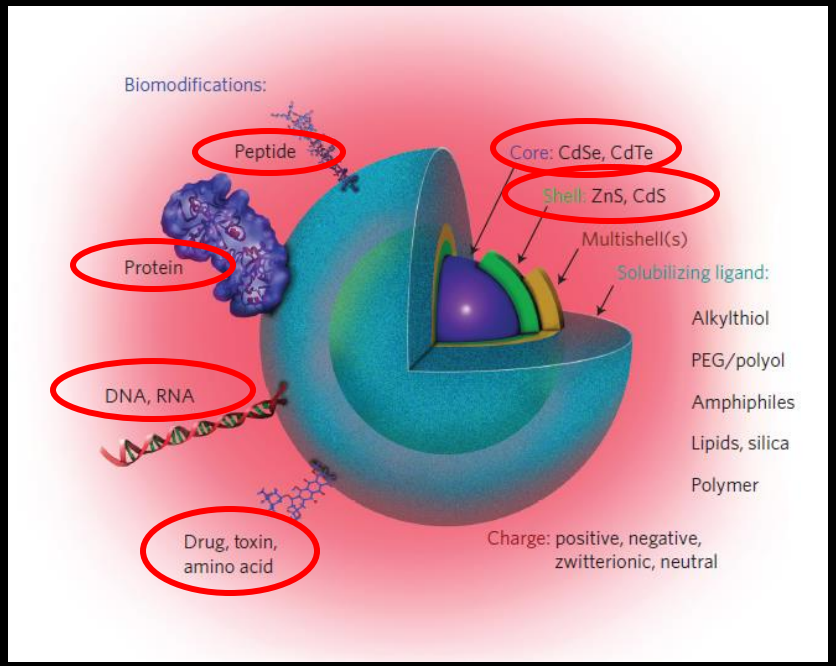
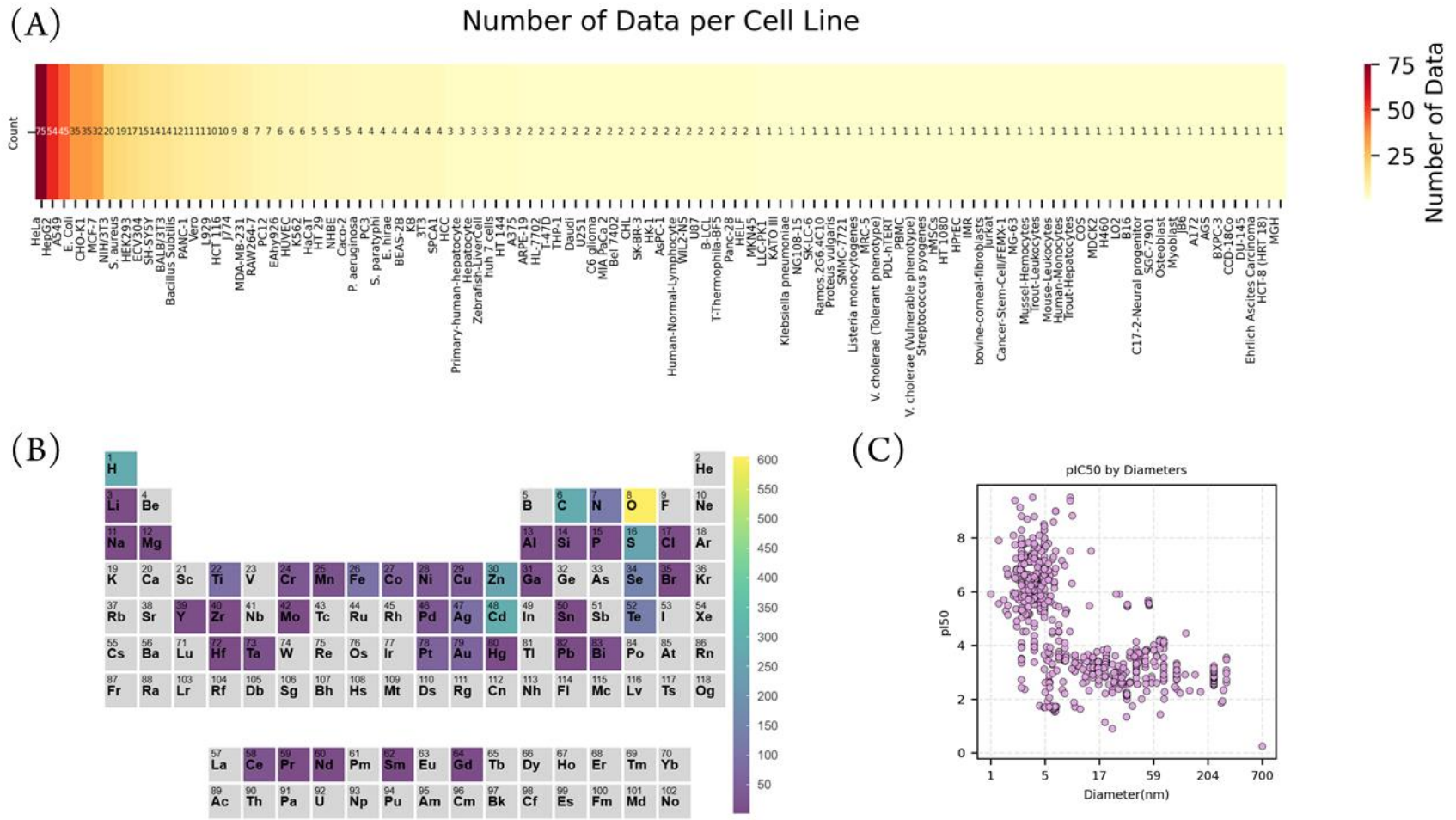


Fig Ref) Oh, E., Liu, R., Nel, A. et al. Meta-analysis of cellular toxicity for cadmium-containing quantum dots. Nature Nanotech 11, 479–486 (2016). https://doi.org/10.1038/nnano.2015.338

Data Preprocessing

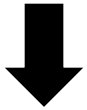


Before	After
921	637

Feature Preparation

Cell Information

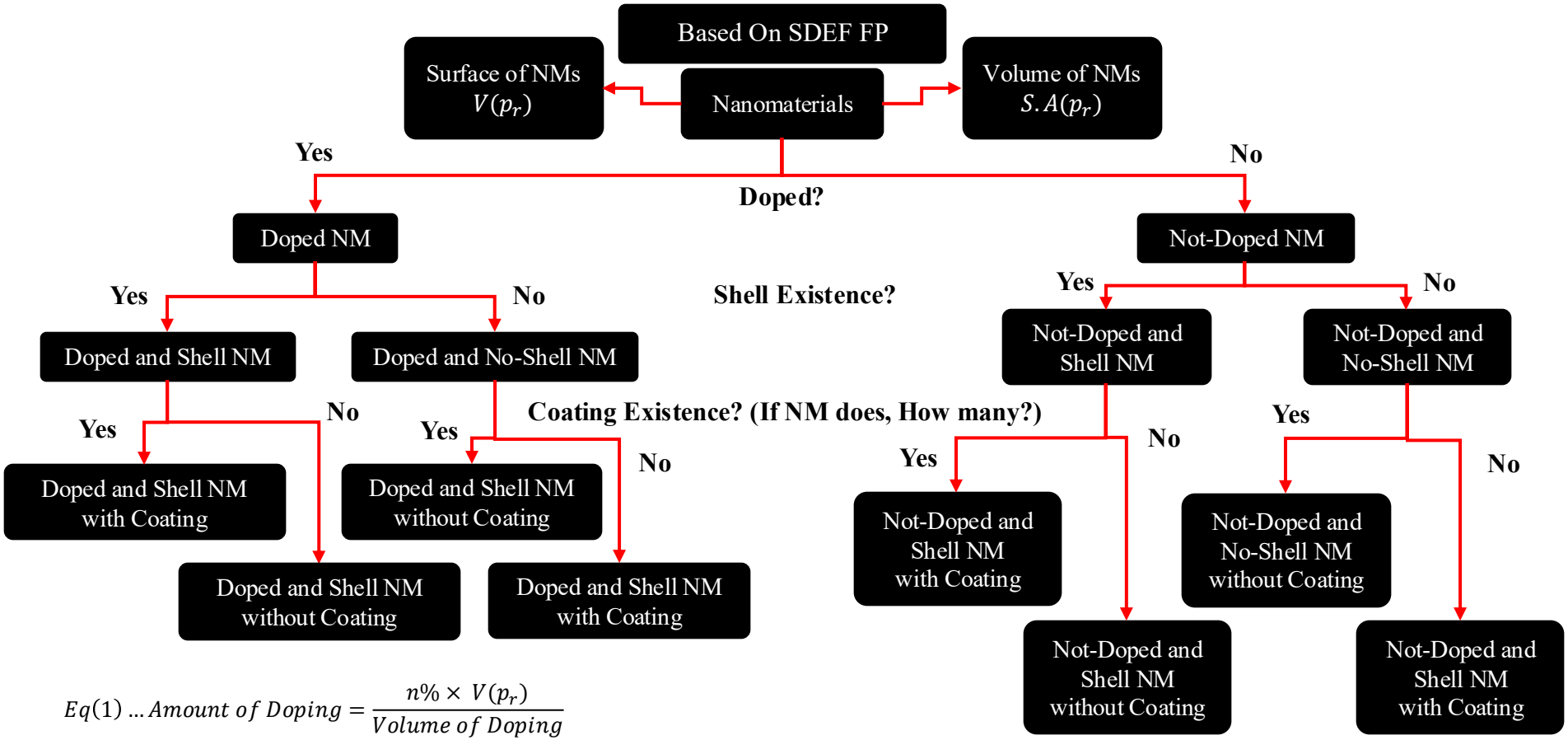
1. Cell Anatomical Type
2. Cell Identification (Cell Type)
3. Cell Source Species
4. Cell Origin
5. Cell Tissue Origin



One-hot Encoded

Feature Preparation

Pipeline (Based on *)



$$Eq(1) \dots Amount\ of\ Doping = \frac{n\% \times V(p_r)}{Volume\ of\ Doping}$$

$$Eq(2) \dots Amount\ of\ Core = \frac{(100-n)\% \times V(p_r)}{Volume\ of\ Core}$$

$$Eq(3) \dots Amount\ of\ Coating = \frac{S.A(p_r) \times 1nm}{Volume\ of\ Coating}$$

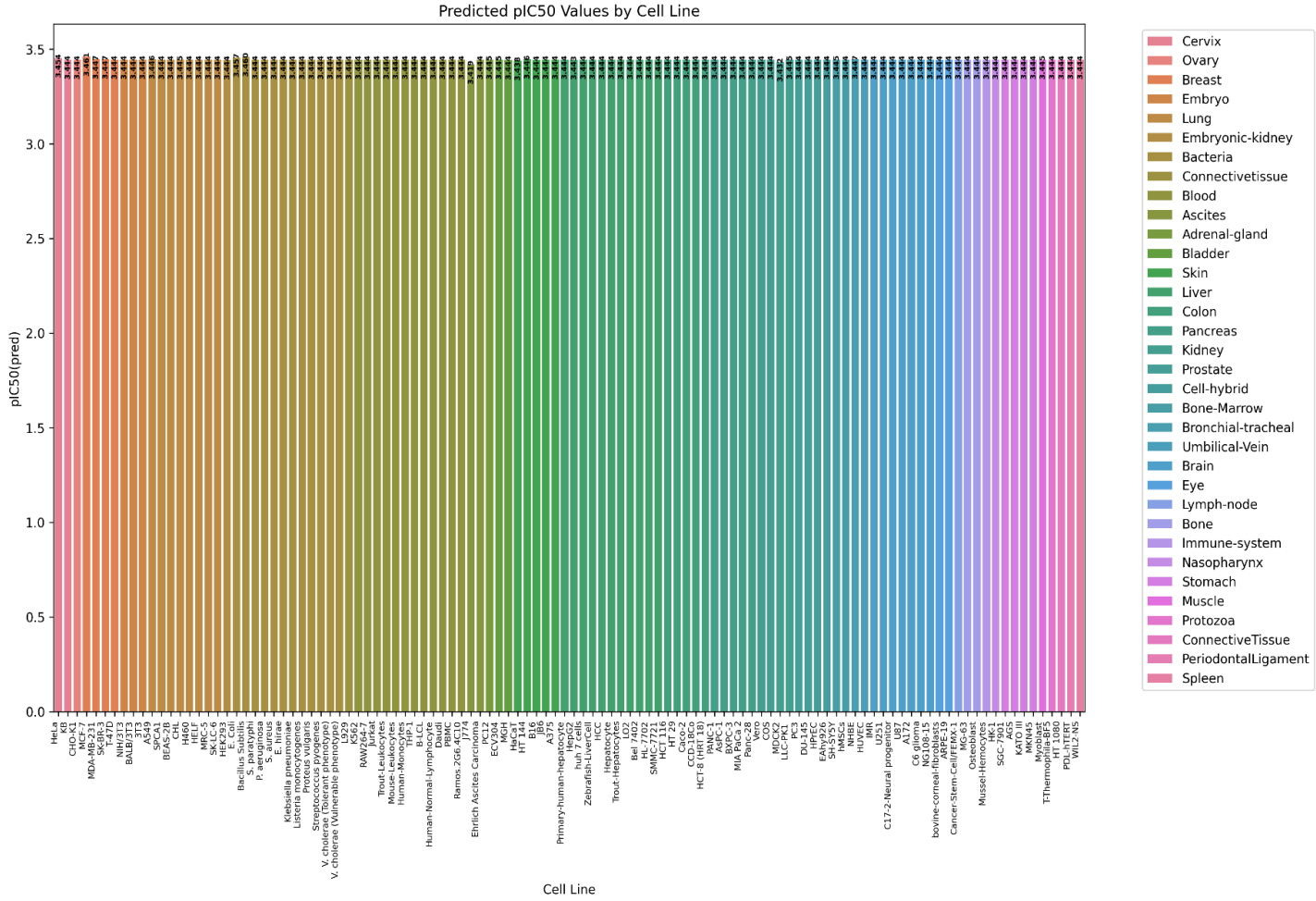
$$Eq(4) \dots Amount\ of\ Shell = \frac{S.A(p_r) \times 1nm}{Volume\ of\ Shell}$$

*Ref) Shin, Hyun Kil et al. "Use of size-dependent electron configuration fingerprint to develop general prediction models for nanomaterials." NanoImpact vol. 21 (2021): 100298. doi:10.1016/j.impact.2021.100298

Development Results

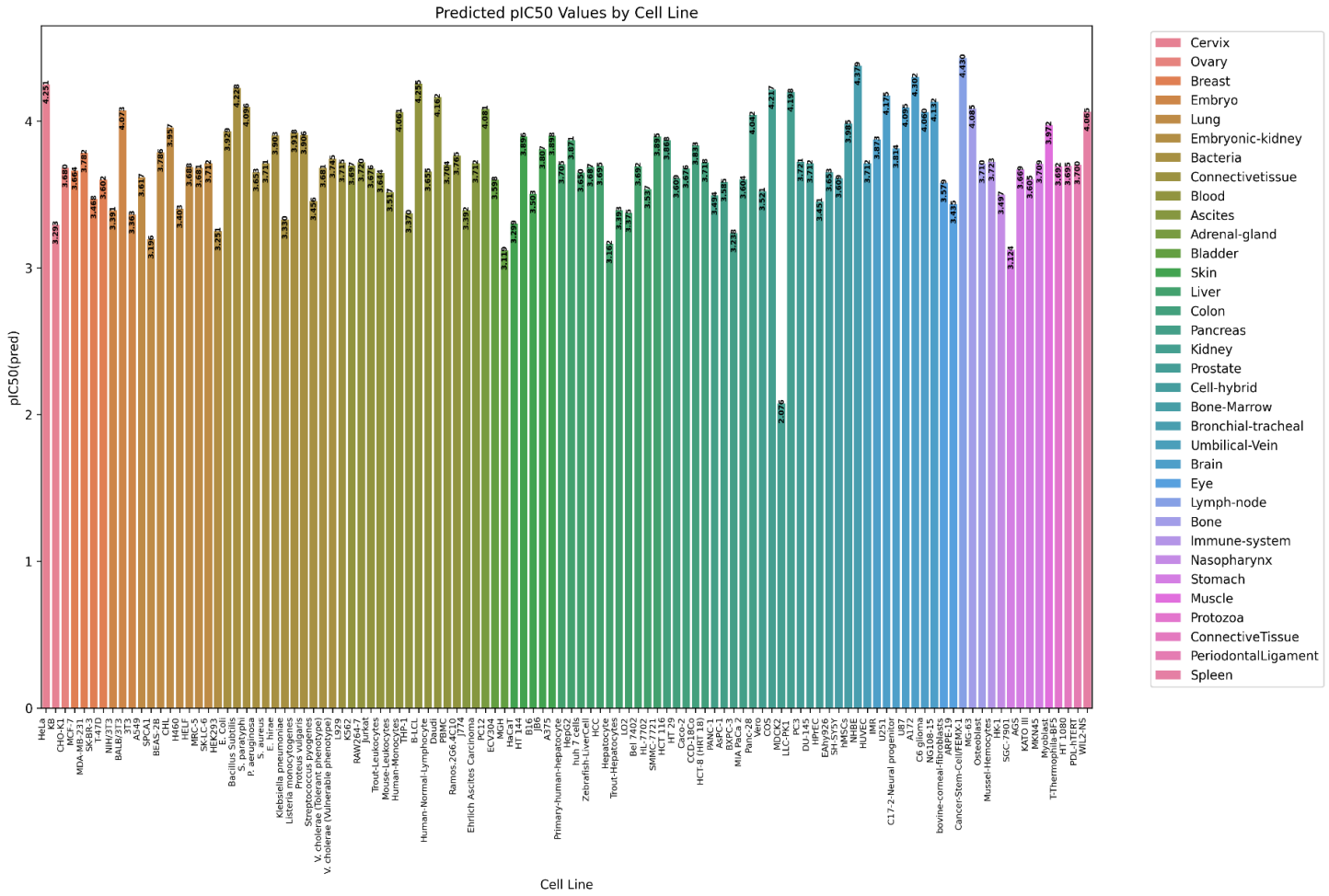
Number of features	Model	$RMSE_{CV}$	R^2_{CV}	$RMSE_{Test}$	R^2_{Test}	$RMSE_{Test}$ over endpoint range (%)	Feature description*
314	CatBoost	0.602 ± 0.013	0.903 ± 0.005	0.623	0.886	6.49%	All cell information & SDEC FP
	ExtraTrees	0.818 ± 0.032	0.820 ± 0.019	0.704	0.855	8.83%	
	SVR	0.691 ± 0.038	0.871 ± 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%	

Development Results



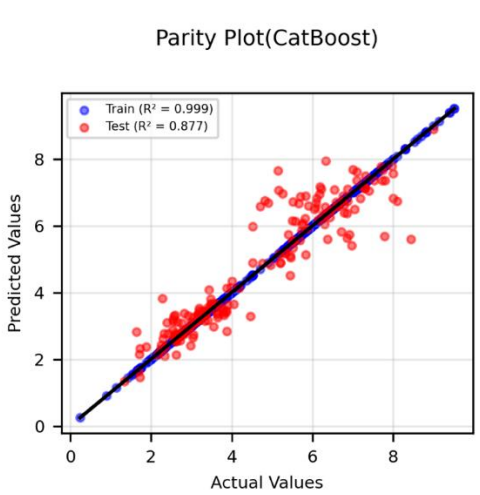
No Distinguished Prediction by Cell Line in SVR Model

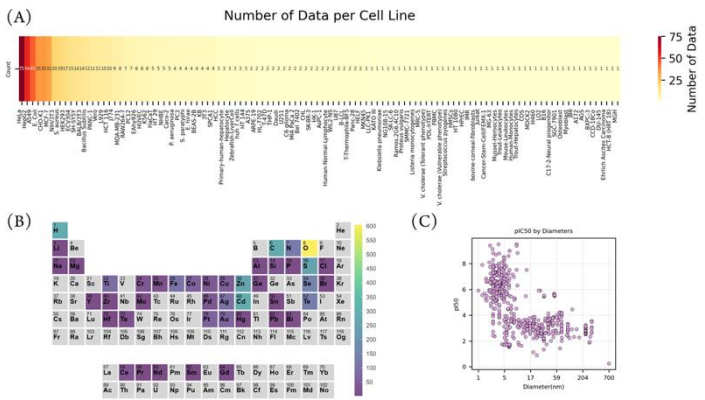
Development Results

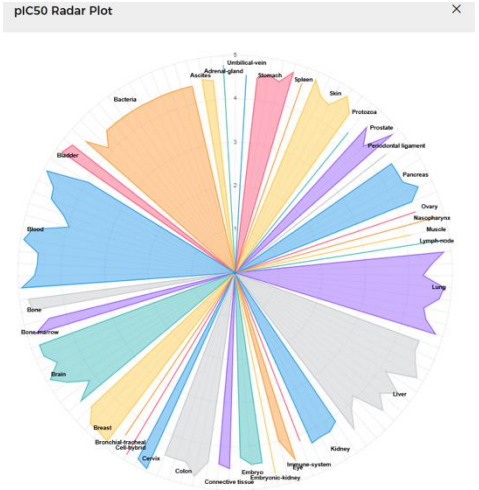


Distinguished Prediction by Cell Line Catboost Model

Development Results



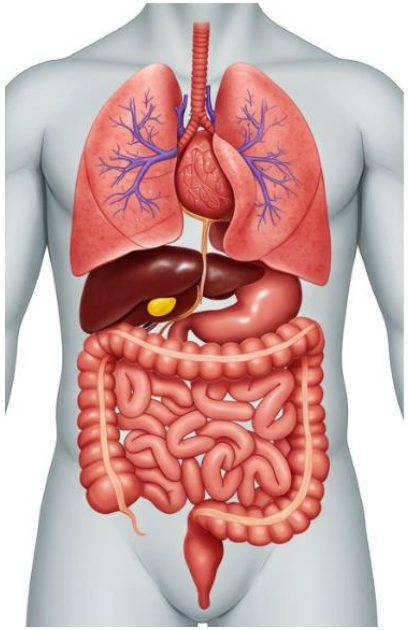


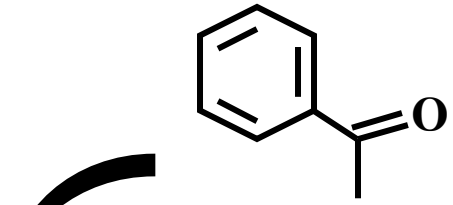


Number of features	Model	$RMSE_{CV}$	R^2_{CV}	$RMSE_{Test}$	R^2_{Test}	$RMSE_{Test}$ over endpoint range (%)	Feature description*
130	CatBoost	0.691 ± 0.029	0.872 ± 0.005	0.649	0.877	7.45%	Cell names & aggregated SDEC FP without spin

1. NanoToxRadar was developed and deployed as a web-based platform for nanotoxicity prediction.
2. Expanded applicability domain of nanotoxicity prediction model by covering MC-NPs over cytotoxicity.
3. The model with SDEC FP and cell one-hot encoded features predicted multi-target cytotoxicity of MC-NPs.

Measuring Drug Consumption by the Gut-Microbiome





Drug

Metabolism

Drug Consumption by Gut Bacteria

Distinct Function, Toxicity

Individual Differences in Drug response

Wouldn't knowing the degree of drug consumption by Gut-Microbiome be the key first step towards personalized medicine?

Measuring Drug Consumption by the Gut-Microbiome

Published in final edited form as:
Nature. 2019 June ; 570(7762): 462–467. doi:10.1038/s41586-019-1291-3.

Mapping human microbiome drug metabolism by gut bacteria and their genes

Michael Zimmermann[*], **Maria Zimmermann-Kogadeeva**[*], **Rebekka Wegmann**[1], and **Andrew L. Goodman**
Department of Microbial Pathogenesis and Microbial Sciences Institute, Yale University School of Medicine, New Haven, CT 06536, USA

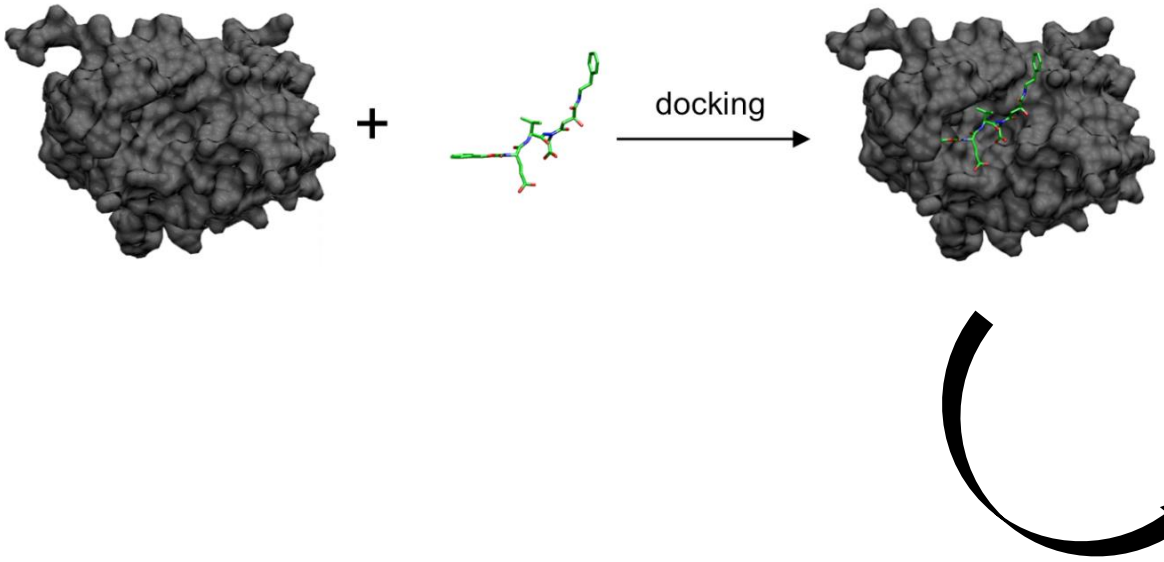
Drugs	Microbiome	Data Points
237	76	3087

Measuring Drug Consumption by the Gut-Microbiome

1. **Collect the Enzymes PDB for the docking.**
   - Microbiome Produced Enzymes which participate the drug metabolism

2. **Docking scores will be features for training the model.**

3. **It worked…?**



Thank you

Q&A

KIT 국가독성과학연구소
Korea Institute of Toxicology

CHAPTER I II III IV

SUPPLEMENTARY

2.3 Model Development

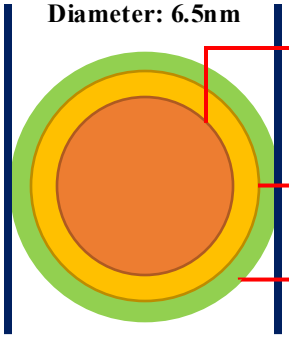
Data Points	Feature
637	314

Main Task

1. To predict endpoint robustly (pIC_{50})
2. To minimize the overfitting issue (1:5 - feature number : data point)*

* Ref) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A., QSAR Modeling: Where Have You Been? Where Are You Going To? Journal of Medicinal Chemistry 2014, 57 (12), 4977-5010.

2.2 Feature Preparation



Not-Doped and Shell NM with Coating

Core: CdSe
Doping: No

Shell: CdS/ZnS

Coating: PEI/PEG (C2H5N/C2H5O2)

$$Eq(2) \dots Amount\ of\ Core = \frac{(100-0)\% \times V(p_r)}{1 \times V(Cd^{+2}) + 1 \times V(Se^{-2})}$$

$$Eq(3) \dots Amount\ of\ Coating = \frac{S.A(p_r) \times 1nm}{2 \times V(C^0) + 5 \times V(H^5) + 1 \times V(N^0) + 2 \times V(C^0) + 5 \times V(H^5) + 2 \times V(O^0)}$$

$$Eq(4) \dots Amount\ of\ Shell = \frac{S.A(p_r) \times 1nm}{1 \times V(Cd^{+2}) + 1 \times V(S^{-2}) + 1 \times V(Zn^{+2}) + 1 \times V(S^{-2})}$$

$$V(p_r) = \frac{4}{3}(6.5nm)^3$$

$$S.A(p_r) = 4\,\pi(6.5nm)^2$$

Core
Amount of Cd = Amount of Core x 1
Amount of Se = Amount of Core x 1

Shell
Amount of Cd = Amount of Shell x 1
Amount of S = Amount of Shell x 1
Amount of Zn = Amount of Shell x 1
Amount of S = Amount of Shell x 1

Coating
Amount of C = Amount of Coating x 2
Amount of H = Amount of Coating x 5
Amount of N = Amount of Coating x 1
Amount of C = Amount of Coating x 2
Amount of H = Amount of Coating x 5
Amount of N = Amount of Coating x 2

atom	1s+	1s-	2s+	2s-	2pz+	2px+	2py+	2pz-	2px-	2py-	3s+	3s-	3pz+	3px+
H	1	0	0	0	0	0	0	0	0	0	0	0	0	0
He	1	-1	0	0	0	0	0	0	0	0	0	0	0	0
Li	1	-1	1	0	0	0	0	0	0	0	0	0	0	0
Be	1	-1	1	-1	0	0	0	0	0	0	0	0	0	0
B	1	-1	1	-1	1	0	0	0	0	0	0	0	0	0
C	1	-1	1	-1	1	1	0	0	0	0	0	0	0	0
N	1	-1	1	-1	1	1	1	0	0	0	0	0	0	0
O	1	-1	1	-1	1	1	1	-1	0	0	0	0	0	0
F	1	-1	1	-1	1	1	1	-1	-1	0	0	0	0	0
Ne	1	-1	1	-1	1	1	1	-1	-1	-1	0	0	0	0
Na	1	-1	1	-1	1	1	1	-1	-1	-1	1	0	0	0
Mg	1	-1	1	-1	1	1	1	-1	-1	-1	1	-1	0	0
Al	1	-1	1	-1	1	1	1	-1	-1	-1	1	-1	1	0
Si	1	-1	1	-1	1	1	1	-1	-1	-1	1	-1	1	1

…

…

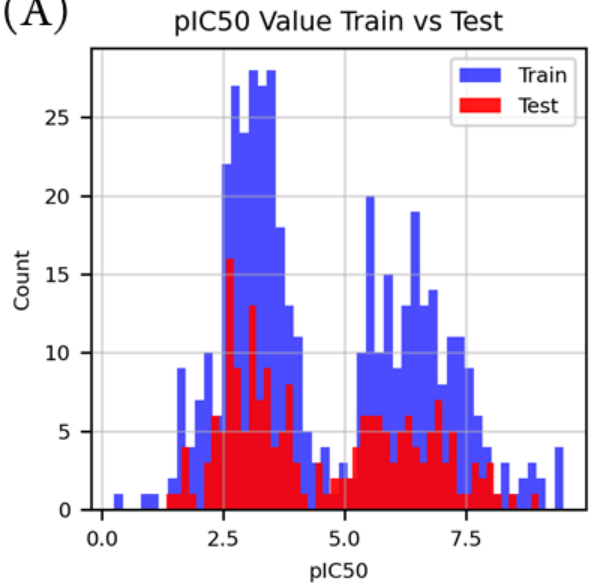
*EC = Electric Configuration

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]
…

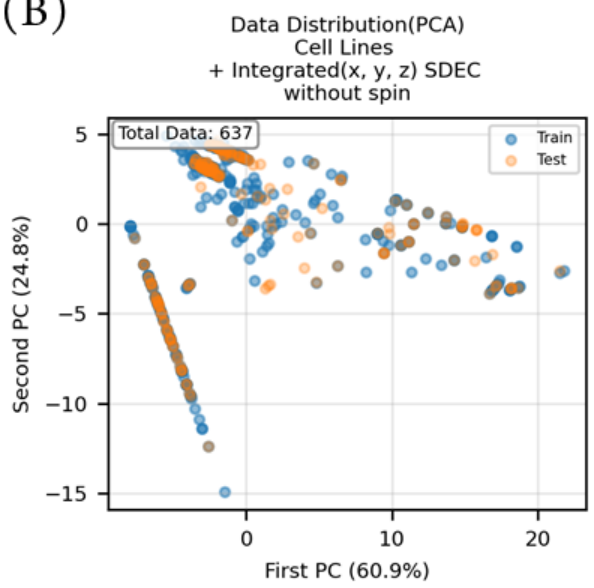
Estimated Electron Configuration Vectors for NM

3.1 Data Visualisation

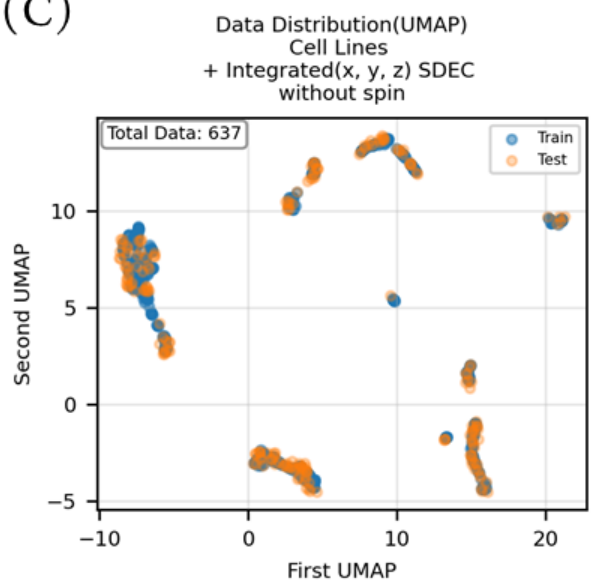
(A)



(B)



(C)



Identical and Independent Distribution (IID)

Lead to robust prediction model

The initial CV by feature descriptions

Model	Number of features	$RMSE_{Train}$	$RMSE_{Validation}$	Comparative index (C.I.)	Feature Description (Number of features)
CatBoost	222	0.244	0.671	2.040	All Cell Information (182) and Integrated(x, y, z) SDEC FP (40)
ExtraTrees		1.009	1.095	0.252	
XGBoost		0.170	0.828	2.546	
GradientBoosting		0.478	0.815	1.323	
RandomForest		0.868	1.023	0.484	
SVR		1.106	1.162	0.154	
Transformer		0.234	0.772	2.234	
MLP		0.420	0.819	1.564	
CatBoost	202	0.265	0.723	1.684	All Cell Information (182) and Integrated(x, y, z) SDEC FP without spin (20)
ExtraTrees		1.147	1.230	0.179	
XGBoost		0.201	0.796	1.990	
GradientBoosting		0.549	0.911	1.058	
RandomForest		0.940	1.077	0.338	
SVR		1.182	1.255	0.156	
Transformer		0.300	0.825	1.693	
MLP		0.459	0.844	1.215	
CatBoost	242	0.281	0.737	2.698	Cell Line (110) and All SDEC FP (132)
ExtraTrees		0.964	1.091	0.506	
XGBoost		0.198	0.847	3.341	
GradientBoosting		0.496	0.853	1.824	
RandomForest		0.809	0.997	0.821	
SVR		1.132	1.198	0.238	
Transformer		0.467	0.835	1.921	
MLP		0.446	0.873	2.131	
CatBoost	150	0.290	0.732	1.157	Cell Line (110) and Integrated(x, y, z) SDEC FP (40)
ExtraTrees		1.023	1.132	0.185	
XGBoost		0.222	0.867	1.425	
GradientBoosting		0.515	0.854	0.761	
RandomForest		0.845	1.019	0.327	
SVR		1.119	1.174	0.090	
Transformer		0.253	0.784	1.298	
MLP		0.406	0.848	0.997	

The initial CV by feature descriptions

Model	Number of features	$RMSE_{Train}$	$RMSE_{Validation}$	Comparative index (C.I.)	Feature Description (Number of features)
CatBoost	276	0.254	0.692	5.230	Cell Line/ Tissue/Organ (144) and All SDEC FP (132)
ExtraTrees		0.966	1.094	0.963	
XGBoost		0.171	0.827	6.557	
GradientBoosting		0.473	0.825	3.530	
RandomForest		0.815	1.003	1.544	
SVR		1.130	1.195	0.455	
Transformer		0.279	0.807	5.407	
MLP		0.480	0.889	3.805	
CatBoost	184	0.258	0.680	1.498	Cell Line/ Tissue/Organ (144) and Integrated(x, y, z) SDEC FP (40)
ExtraTrees		1.013	1.106	0.202	
XGBoost		0.181	0.857	1.906	
GradientBoosting		0.482	0.813	0.982	
RandomForest		0.868	1.034	0.387	
SVR		1.115	1.170	0.115	
Transformer		0.268	0.790	1.597	
MLP		0.432	0.857	1.198	
CatBoost	164	0.283	0.759	1.312	Cell Line/Tissue/Organ (144) and Integrated(x, y, z) SDEC FP without spin (20)
ExtraTrees		1.147	1.224	0.132	
XGBoost		0.224	0.831	1.529	
GradientBoosting		0.553	0.902	0.810	
RandomForest		0.951	1.101	0.286	
SVR		1.197	1.271	0.122	
Transformer		0.296	0.812	1.329	
MLP		0.497	0.946	0.994	
CatBoost	264	0.250	0.708	4.062	Cell Line/Anatomical Type (132) and All SDEC FP (132)
ExtraTrees		0.956	1.089	0.767	
XGBoost		0.181	0.840	4.925	
GradientBoosting		0.472	0.849	2.787	
RandomForest		0.802	0.992	1.202	
SVR		1.129	1.195	0.347	
Transformer		0.290	0.758	3.874	
MLP		0.460	0.813	2.729	

The initial CV by feature descriptions

Model	Number of features	$RMSE_{Train}$	$RMSE_{Validation}$	Comparative index (C.I.)	Feature Description (Number of features)
CatBoost	152	0.282	0.758	1.218	Cell Line/Anatomical Type (132) and Integrated(x, y, z) SDEC FP without spin (20)
ExtraTrees		1.162	1.262	0.153	
XGBoost		0.220	0.803	1.407	
GradientBoosting		0.562	0.927	0.763	
RandomForest		0.948	1.099	0.266	
SVR		1.198	1.274	0.116	
Transformer		0.307	0.830	1.222	
MLP		0.457	0.877	0.928	
CatBoost	130	0.317	0.778	1.011	Cell Line (110) and Integrated(x, y, z) SDEC FP without spin (20)
ExtraTrees		1.160	1.247	0.120	
XGBoost		0.252	0.856	1.204	
GradientBoosting		0.560	0.917	0.664	
RandomForest		0.946	1.106	0.248	
SVR		1.205	1.278	0.098	
Transformer		0.261	0.805	1.154	
MLP		0.499	0.933	0.795	
CatBoost	172	0.264	0.708	1.386	Cell Line/Anatomical Type (132) and Integrated(x, y, z) SDEC FP (40)
ExtraTrees		1.016	1.122	0.209	
XGBoost		0.196	0.851	1.703	
GradientBoosting		0.483	0.844	0.945	
RandomForest		0.843	1.024	0.391	
SVR		1.115	1.171	0.105	
Transformer		0.306	0.756	1.317	
MLP		0.427	0.826	1.069	

Deep Learning Architecture

MLP (Multi Layer Perceptron)

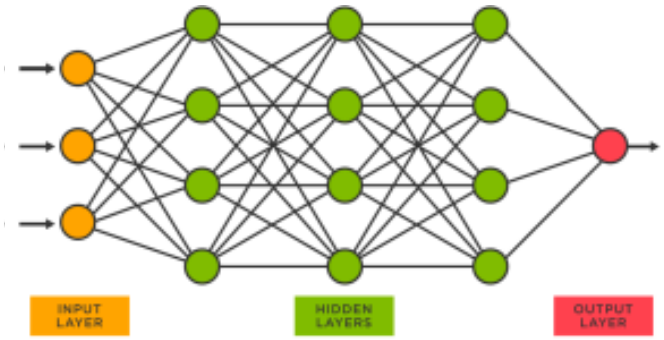


Fig Ref) https://vitalflux.com/sklearn-neural-network-regression-example-mlpregressor/

Transformer Encoder (Modified)