Dataset Curation, Assessment of their Quality, and Prediction Model Developments for Safe and Sustainable Nanotechnology (S2NANO)

Prof. Tae Hyun Yoon, CEO/Ph.D.

Hanyang University & Yoon Idea Lab. Co. Ltd.

References - S²NANO:PredictNano

J.S. Choi, T.X. Trinh, T. H. Yoon, J. Kim, H. G. Byun* (2019)* "*Quasi-QSAR for predicting the cell viability of human lung and skin cells exposed to different metal oxide nanomaterials* " ***Chemosphere****, 217, 243*

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

*M.K. Ha, T.X. Trinh, J.S. Choi, D. Maulina, H.G. Byun, T.H.Yoon**, (2018) ***"Toxicity Classification of Oxide Nanomaterials: Effects of Data Gap Filling and PChem Score-based Screening Approaches."*** ***Scientific Reports*****, 8**, 3141

*T.X. Trinh, M.K. Ha, J.S. Choi, H.G. Byun, T.H. Yoon** (2018) ***"Dataset Curation, Assessment of their Quality and Completeness, and nanoSAR Classification Model Development for Metallic Nanoparticles"*** ***Environmental Science: Nano*** 5, 1902-1910

T.X. Trinh, J.S. Choi, H. Jeon, H.G. Byun, T.H.Yoon, J.Kim**,(2018) "Quasi-SMILES based Nano-QSAR model to predict the cytotoxicity of multi-walled carbon nanotubes to human lung cells." ***Chemical Research in Toxicolology,***31(3), 183

Sunil Kr Jha, TH Yoon, Zhaoqing Pan* (2018) "Multivariate statistical analysis for selecting optimal descriptors in the toxicity modeling of nanomaterials", ***Computers in Biology and Medicine***, 99, 161

D.W. Boukhvalov, T.H.Yoon*, (2017) "Development of Theoretical Descriptors for Cytotoxicity Evaluation of Metallic Nanoparticles." ***Chemical Research in Toxicolology***, 30 (8), 1549

How can we bridge the gap between
Excellence in nano R&D and Profitable nano Industry ?

Excellence in nano R&D

Profitable nano Industry

Nanosafety ?

Safe & Sustainable Nanotechnology Portal :
A Toolbox for Nanomaterial Characterization & Safety Assessments
www.s2nano.org

Safe NANO Sustainable

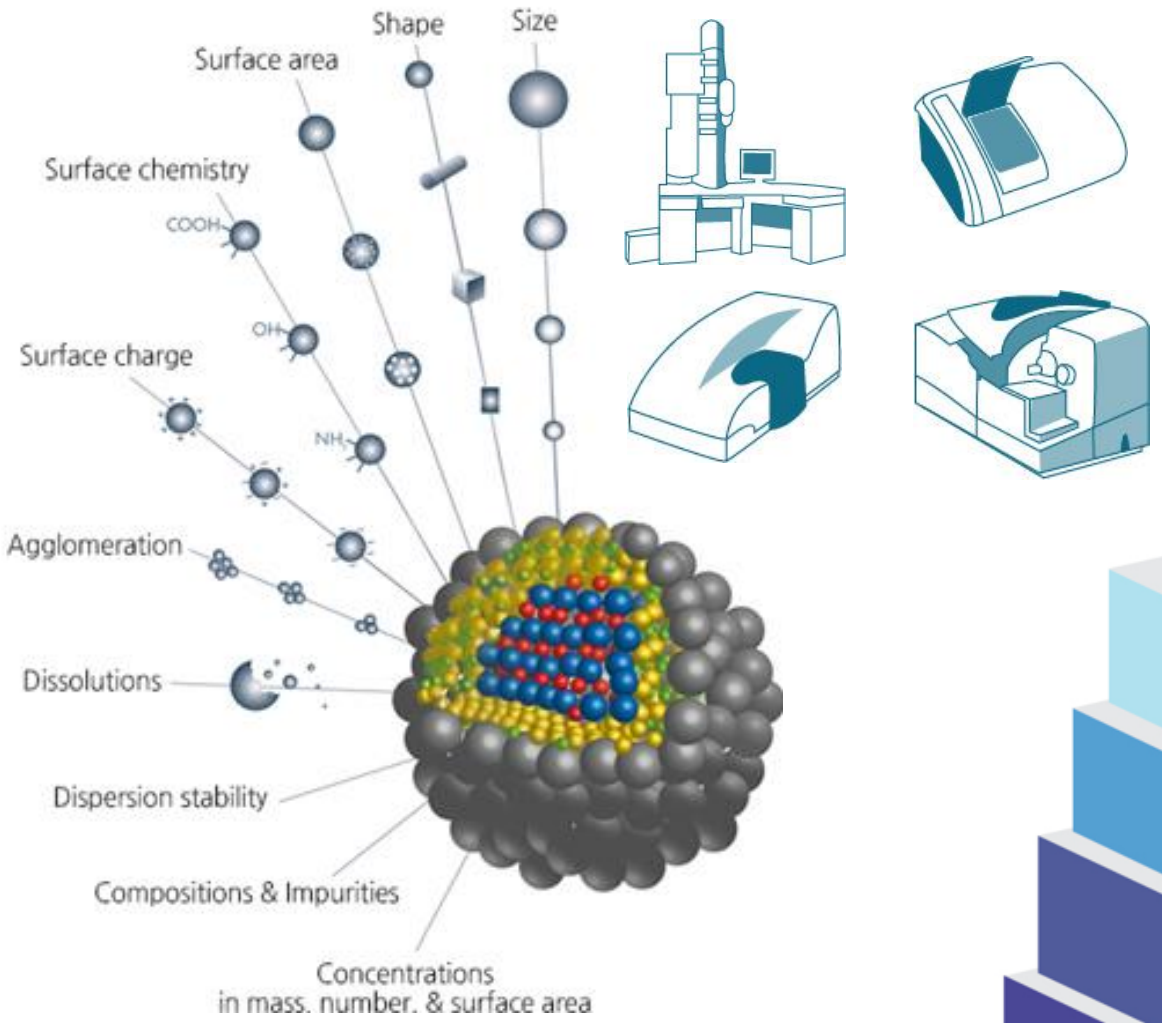
http://portal.s2nano.org

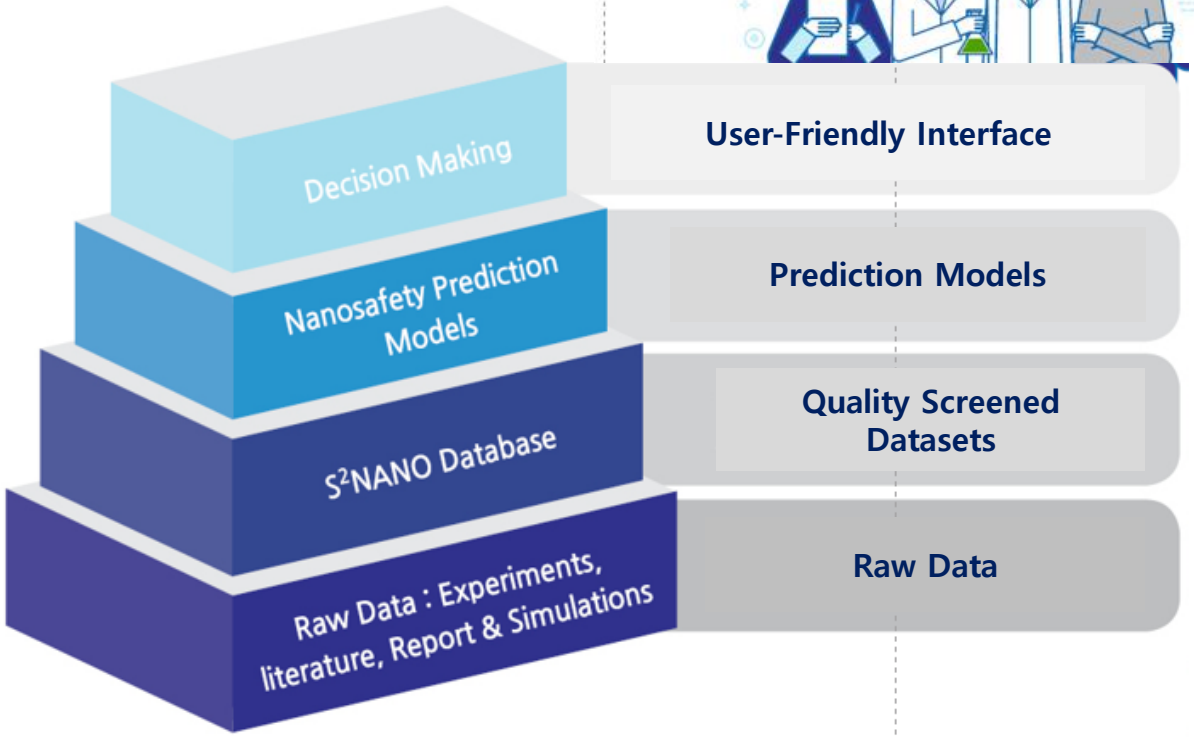
"사용자 친화적 나노안전성 예측시스템"

2013 2014 2015 2016 2017 2018 2019 2020 2021 2022

S2NANO : PredictNANO
2012.12 - 2017.11
종료시 TRL 5단계

Measurements to Models For Nanomaterials



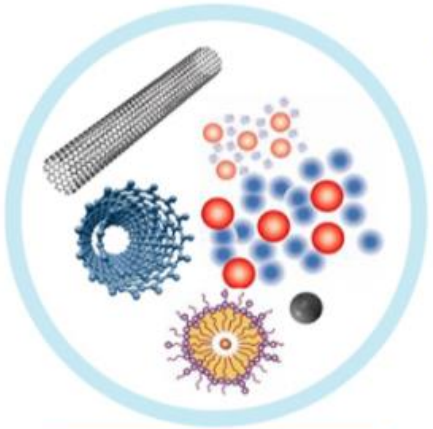


User Friendly Nanosafety Prediction System

Input

구성성분	SiO_2
입자크기	3nm ~300nm
표면저하	-30mV ~ +30mV
입자모양	구형

Safety?



Manufactured Nanomaterials

1 | Reference NM Library

2 | Standard NM Library

3 | Manufactured NM Library

PChem Data (Measured)

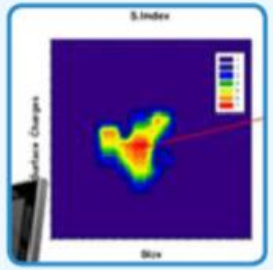
PChem Data (Literature)

NanoToxicity Prediction Models

Tox Data (Measured)

Tox Data (Literature)

Output



높은 안전성 지수 값을 가지는 입자크기와 표면전하 영역

Safe or Unsafe

Safety Index

Characteristics of Nanosafety Data ?

Small, Unbalanced, & Heterogeneous datasets with many missing values

How to overcome these complexities of data ?

Comprehensive database with physicochemical & toxicity data of nanomaterials

Dataset curation based on the assessment of data quality / completeness

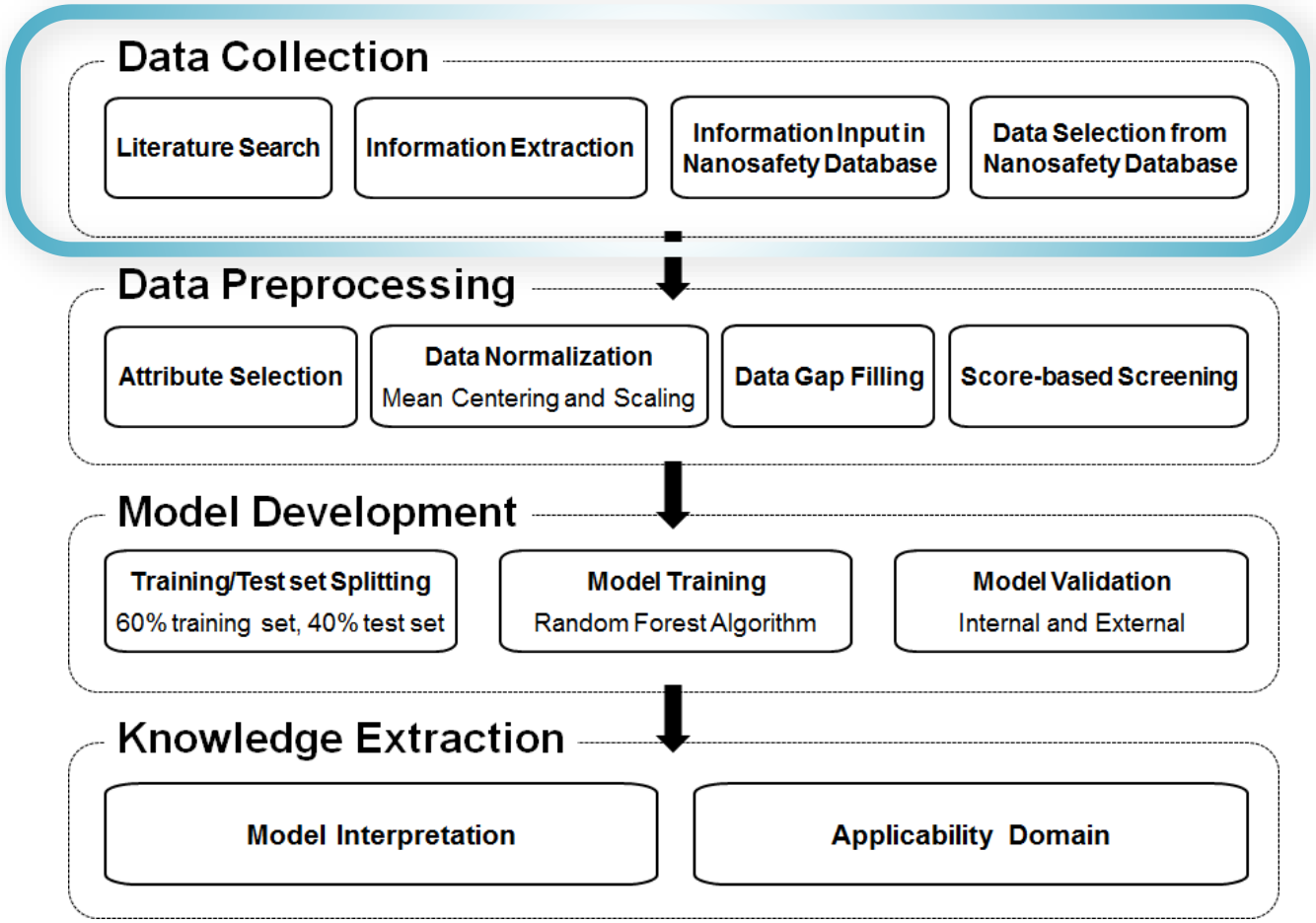
Development of generalized prediction models with wider applicability domains

Model Development Workflow in S²NANO

Core Dataset from our Own Experiments

Extended Dataset from Literature Mining

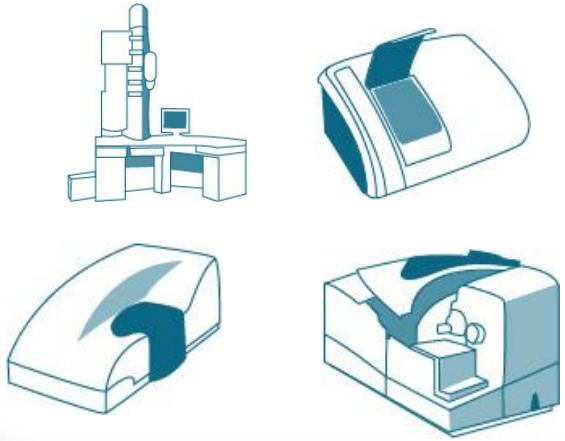
- Info.DB, Mat.DB, QM DB,
- PChem. DB, Tox DB (in vitro, in vivo, eco).,

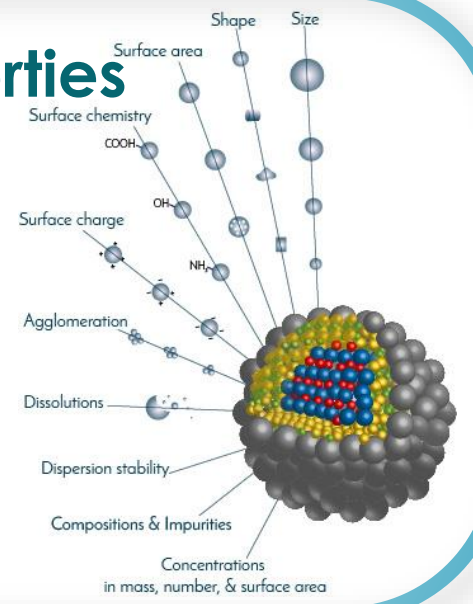


Experimental : PChem

TEM / SEM
DLS / NTA
ICP-MS
Raman / Infrared
STXM

Measurement of Properties





Core Dataset
from our Own Experiments
- Core : PChem DB

Nanosafety database

S2afe NANO Sustainable

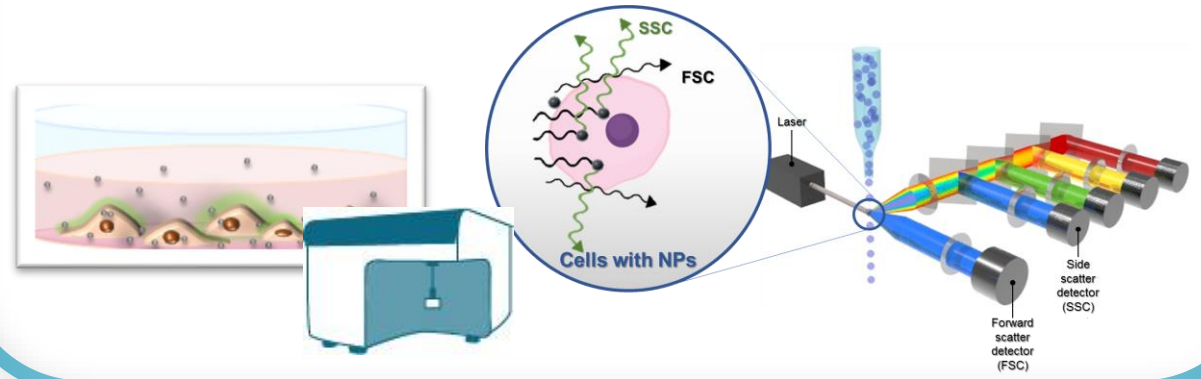
www.s2nano.org

Data Collection - Experimental

Experimental : in vitro Tox

MTT/ MTS
CCK-8
CellTiter-Glo

Assessment of in vitro toxicity



Core Dataset
from our Own Experiments
- Core : Tox(in vitro) DB

Nanosafety database

S2afe NANO Sustainable

www.s2nano.org

Data Collection - Literature Mining

Literature Search

PubMed

SciFinder®

Google Scholar

Selection of Articles on Nanosafety

Extraction of Nanoinformation

: Info/Mat/Pchem/Tox

Extended dataset

from Literature Mining

- Info.DB, Mat.DB, QM DB,
- PChem DB, Tox DB (in vitro, in vivo, eco).,

Nanosafety database

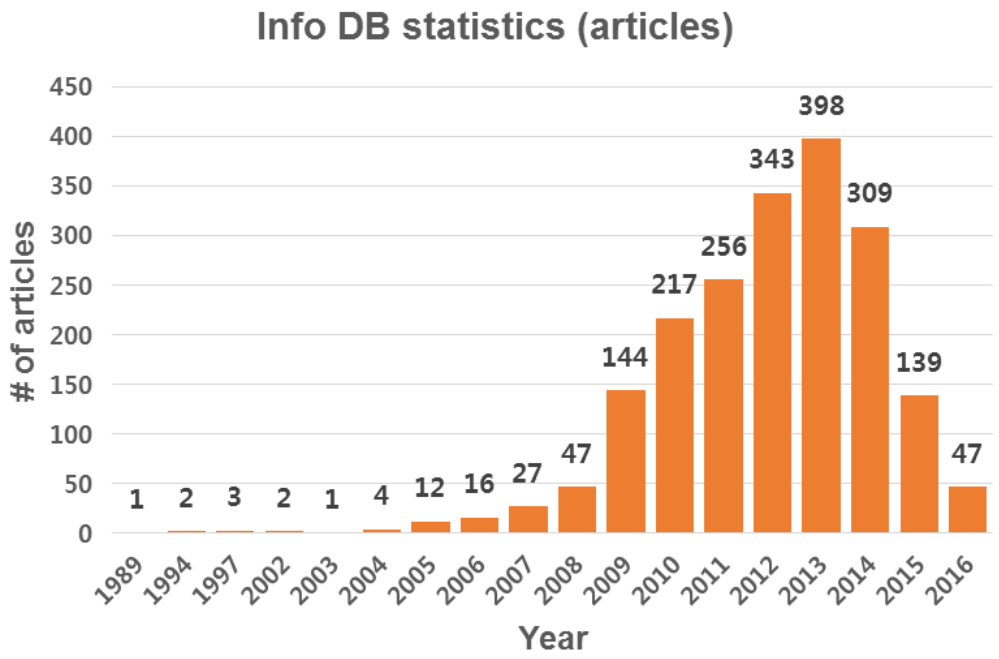
Safe NANO Sustainable

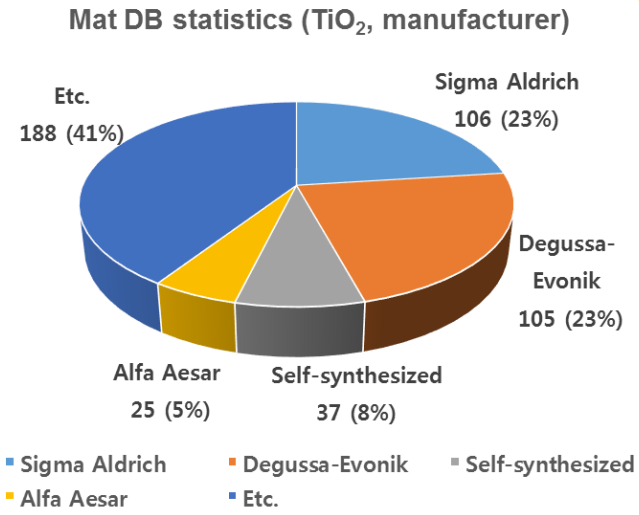
www.s2nano.org

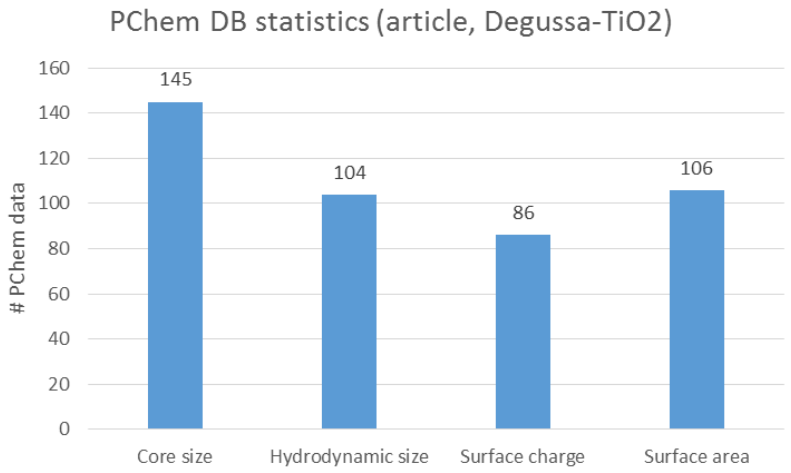
Core & Extended dataset

from Experiment & Literature Mining

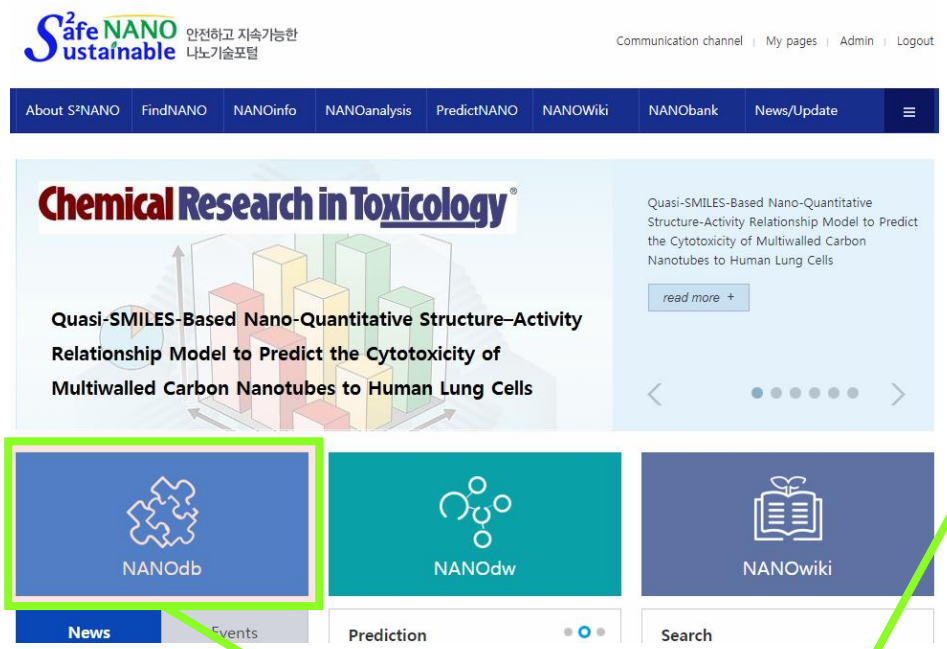
- Info.DB, Mat.DB, QM DB,
- PChem DB, Tox DB (in vitro, in vivo, eco).,

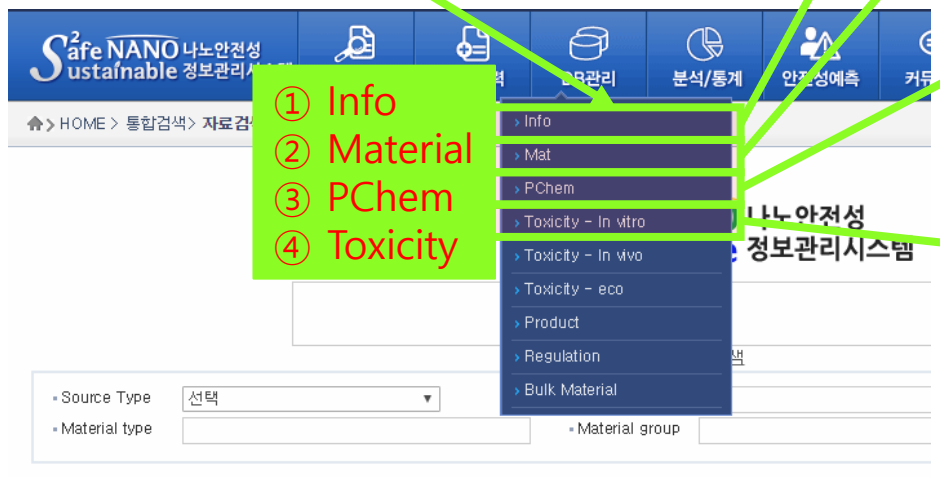




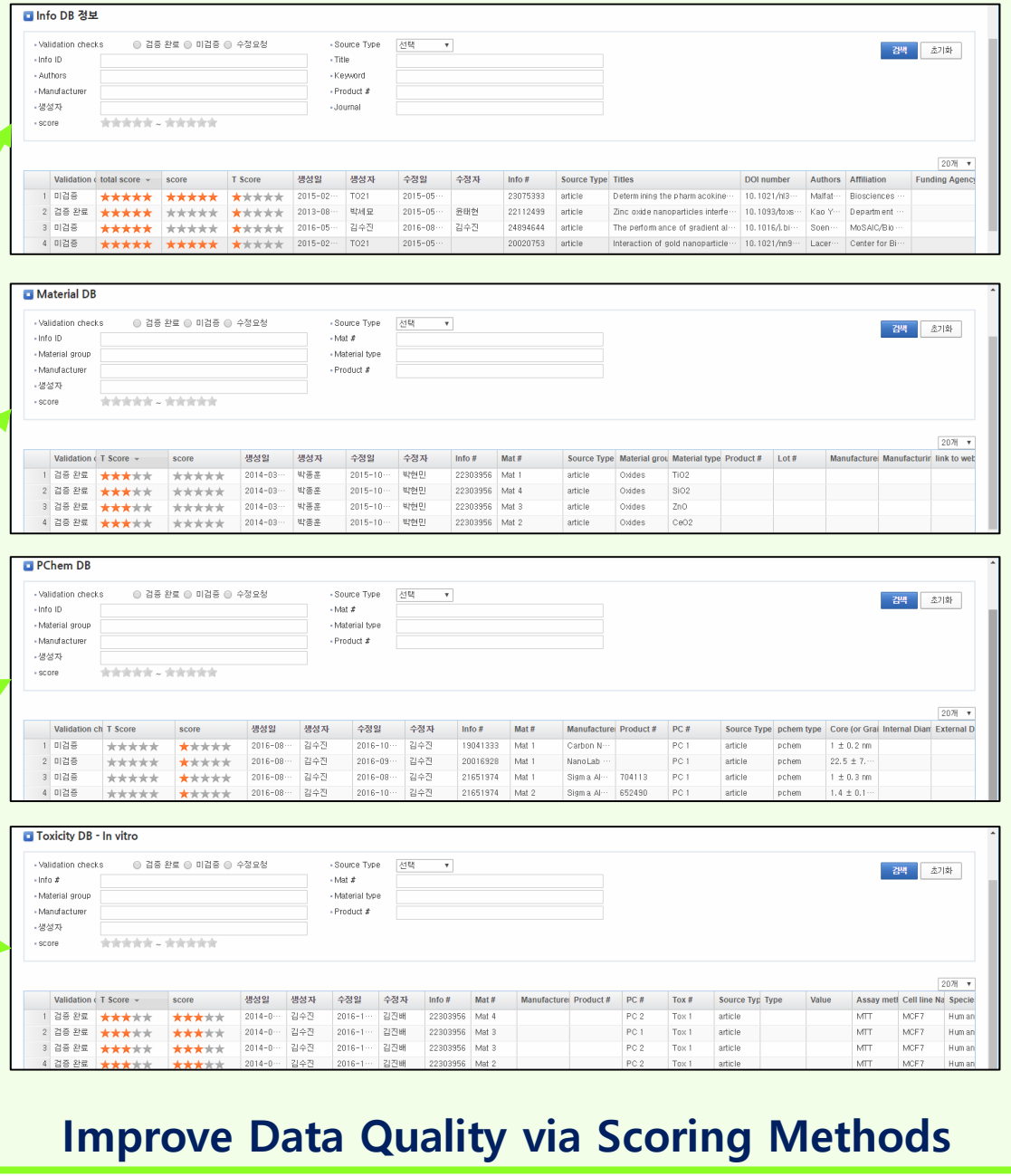


Collected Database



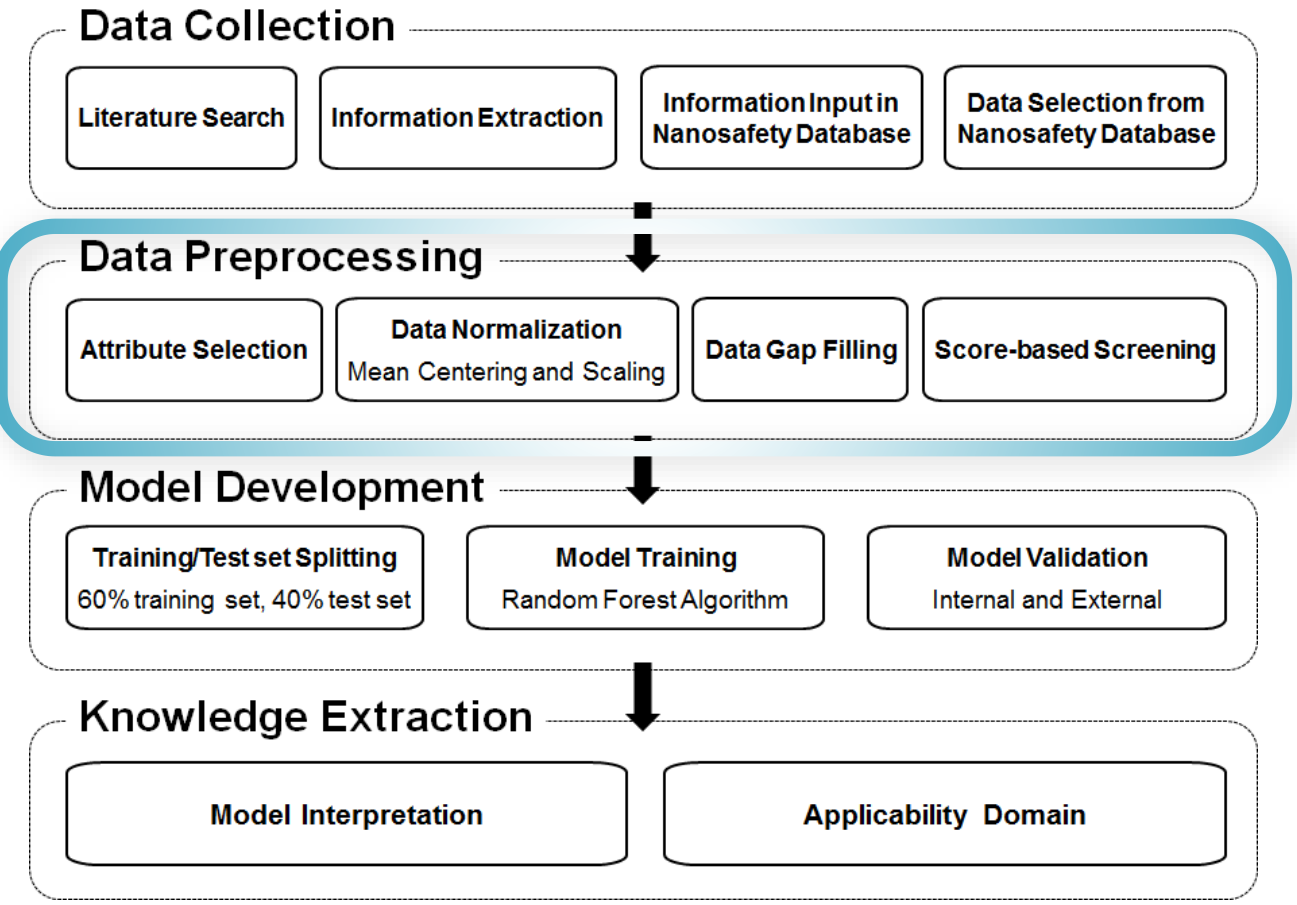


① Info
② Material
③ PChem
④ Toxicity



Improve Data Quality via Scoring Methods

Model Development Workflow in S²NANO

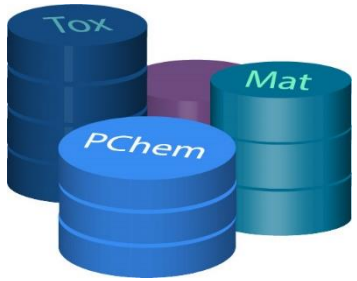


Data Completeness (Missing Data Problem)

Data Quality (Heterogeneous Source of Data)

Data Imbalance (NonToxic >> Toxic)

Data Preprocessing - Attribute Selection



Nanosafety database

Safe NANO Sustainable

www.s2nano.org

Dose

PChem

- Core Size
- Hydrodynamic Size
- Surface Charge
- Surface Area
- Measurement Methods for Each PChem Attributes

In vitro Tox

- Assay
- Type/Name/ Species/Origin of Cell-line
- Exposure time
- Cell Viability

QM

- Formation enthalpy ΔH_{sf} (eV)
- Conduction band energy E_c (eV)
- Valence band energy E_v (eV)
- Electronegativity χ_{MeO} (eV)

Data Preprocessing - Original Dataset (Dataset I)

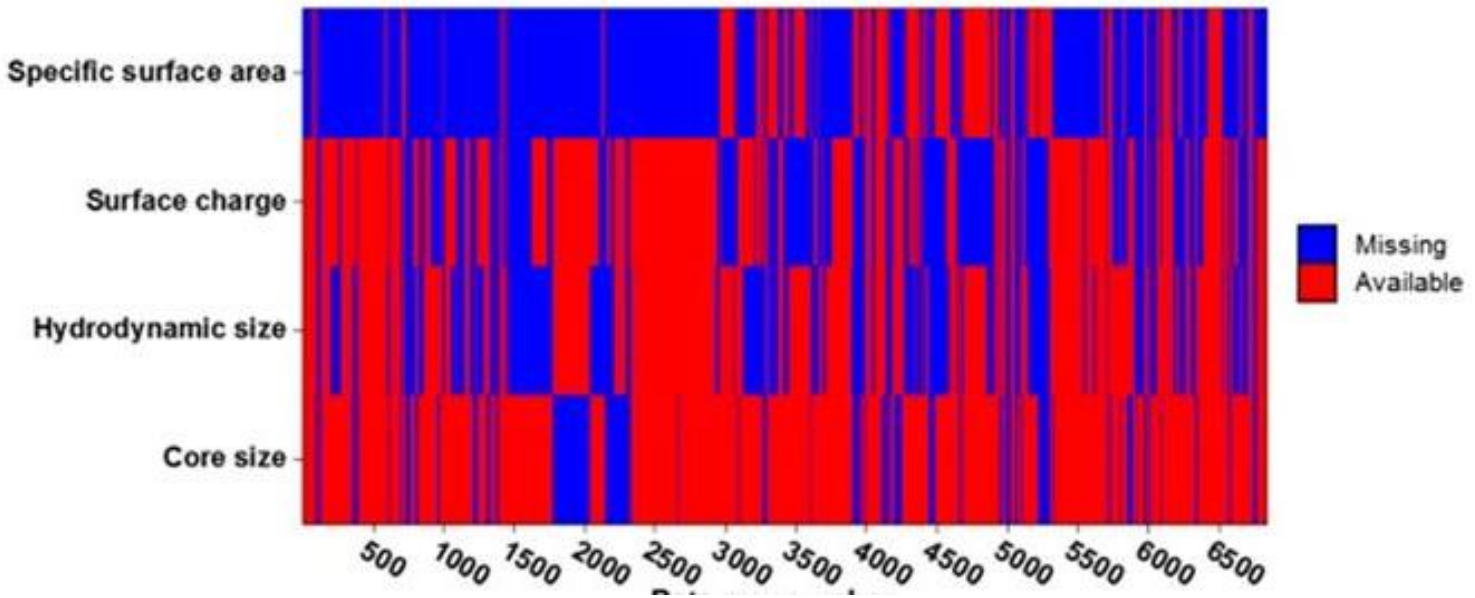
- Total 20 attributes, but only 14 attributes were used as Descriptors
  - Dose(1) / Pchem(8) / Tox(7) / QM(4) attributes
  - Measurement Methods attributes were not used for Model development (-4)
  - Cell name attribute was not used for Model development (-1)
  - Cell Viability was used for Toxic/Non-Toxic Endpoint (-1)

- Toxic/Non-Toxic as Endoint
  - Toxic when Cell Viability < 50 %
  - NonToxic when Cell Viability ≥ 50 %

- 216 articles selected from ~600 pdf files
- 26 oxide NPs
- 6,842 data rows

	A	B	F	G	H	I	J	K	L	M	N	O	P	Q
1	Pubmed ID	Material typ	Method core size	Hydro size (nm)	Method hydro size	Surface charge (mV)	Method surface charge	Surface area (m2/g)	Method surface area	ΔHsf (eV)	Ec (eV)	Ev (eV)	χMeO (eV)	Assay
2	22502734	Al2O3	TEM	260.4	DLS	0	Zeta-potential			-17.345	-1.51	-9.81	5.67	MTS
3	22502734	Al2O3	TEM	260.4	DLS	0	Zeta-potential			-17.345	-1.51	-9.81	5.67	MTS
4	22502734	Al2O3	TEM	260.4	DLS	0	Zeta-potential			-17.345	-1.51	-9.81	5.67	MTS
5	22502734	Al2O3	TEM	260.4	DLS	0	Zeta-potential			-17.345	-1.51	-9.81	5.67	MTS
6	22502734	Al2O3	TEM	260.4	DLS	0	Zeta-potential			-17.345	-1.51	-9.81	5.67	MTS
7	22502734	Al2O3	TEM	260.4	DLS	0	Zeta-potential			-17.345	-1.51	-9.81	5.67	MTS
8	22502734	Al2O3	TEM	260.4	DLS	0	Zeta-potential			-17.345	-1.51	-9.81	5.67	MTS
9	22502734	Al2O3	TEM	260.4	DLS	0	Zeta-potential			-17.345	-1.51	-9.81	5.67	MTS
10	22502734	Al2O3	TEM	260.4	DLS	0	Zeta-potential			-17.345	-1.51	-9.81	5.67	MTS
11	22502734	Al2O3	TEM	260.4	DLS	0	Zeta-potential			-17.345	-1.51	-9.81	5.67	MTS
12	22502734	Al2O3	TEM	260.4	DLS	0	Zeta-potential			-17.345	-1.51	-9.81	5.67	ATP
13	22502734	Al2O3	TEM	260.4	DLS	0	Zeta-potential			-17.345	-1.51	-9.81	5.67	ATP
6813	22502734	ZrO2	TEM	312.3	DLS	-12.8	Zeta-potential			-11.252	-3.19	-8.23	5.62	ATP
6814	22502734	ZrO2	TEM	312.3	DLS	-12.8	Zeta-potential			-11.252	-3.19	-8.23	5.62	ATP
6815	22502734	ZrO2	TEM	312.3	DLS	-12.8	Zeta-potential			-11.252	-3.19	-8.23	5.62	ATP
6816	22502734	ZrO2	TEM	312.3	DLS	-12.8	Zeta-potential			-11.252	-3.19	-8.23	5.62	ATP
6817	22502734	ZrO2	TEM	312.3	DLS	-12.8	Zeta-potential			-11.252	-3.19	-8.23	5.62	ATP
6818	22502734	ZrO2	TEM	312.3	DLS	-12.8	Zeta-potential			-11.252	-3.19	-8.23	5.62	ATP
6819	22502734	ZrO2	TEM	312.3	DLS	-12.8	Zeta-potential			-11.252	-3.19	-8.23	5.62	ATP
6820	22502734	ZrO2	TEM	312.3	DLS	-12.8	Zeta-potential			-11.252	-3.19	-8.23	5.62	ATP
6821	22502734	ZrO2	TEM	312.3	DLS	-12.8	Zeta-potential			-11.252	-3.19	-8.23	5.62	LDH
6822	22502734	ZrO2	TEM	312.3	DLS	-12.8	Zeta-potential			-11.252	-3.19	-8.23	5.62	LDH
6823	22502734	ZrO2	TEM	312.3	DLS	-12.8	Zeta-potential			-11.252	-3.19	-8.23	5.62	LDH
6824	22502734	ZrO2	TEM	312.3	DLS	-12.8	Zeta-potential			-11.252	-3.19	-8.23	5.62	LDH
6825	22502734	ZrO2	TEM	312.3	DLS	-12.8	Zeta-potential			-11.252	-3.19	-8.23	5.62	LDH
6826	22502734	ZrO2	TEM	312.3	DLS	-12.8	Zeta-potential			-11.252	-3.19	-8.23	5.62	LDH
6827	22502734	ZrO2	TEM	312.3	DLS	-12.8	Zeta-potential			-11.252	-3.19	-8.23	5.62	LDH
6828	22502734	ZrO2	TEM	312.3	DLS	-12.8	Zeta-potential			-11.252	-3.19	-8.23	5.62	LDH
6829	22502734	ZrO2	TEM	312.3	DLS	-12.8	Zeta-potential			-11.252	-3.19	-8.23	5.62	LDH
6830	22502734	ZrO2	TEM	312.3	DLS	-12.8	Zeta-potential			-11.252	-3.19	-8.23	5.62	LDH
6831	24983896	ZrO2	TEM	661.4	DLS	-8.5	Zeta-potential			-27.67	-1.03	-10.92	4.95	CytoTox-Glo
6832	24983896	ZrO2	TEM	661.4	DLS	-8.5	Zeta-potential			-27.67	-1.03	-10.92	4.95	CytoTox-Glo
6833	24983896	ZrO2	TEM	661.4	DLS	-8.5	Zeta-potential			-27.67	-1.03	-10.92	4.95	CytoTox-Glo
6834	24983896	ZrO2	TEM	661.4	DLS	-8.5	Zeta-potential			-27.67	-1.03	-10.92	4.95	CytoTox-Glo
6835	24983896	ZrO2	TEM	661.4	DLS	-8.5	Zeta-potential			-27.67	-1.03	-10.92	4.95	CytoTox-Glo
6836	24983896	ZrO2	TEM	661.4	DLS	-8.5	Zeta-potential			-27.67	-1.03	-10.92	4.95	CytoTox-Glo
6837	24983896	ZrO2	TEM	661.4	DLS	-8.5	Zeta-potential			-27.67	-1.03	-10.92	4.95	CytoTox-Glo
6838	24983896	ZrO2	TEM	661.4	DLS	-8.5	Zeta-potential			-27.67	-1.03	-10.92	4.95	CytoTox-Glo
6839	21800953	ZrO2		505	DLS	-23	Zeta-potential			-11.252	-3.19	-8.23	5.62	MTT
6840	21800953	ZrO2		505	DLS	-23	Zeta-potential			-11.252	-3.19	-8.23	5.62	MTT
6841	21800953	ZrO2		505	DLS	-23	Zeta-potential			-11.252	-3.19	-8.23	5.62	MTT
6842	21800953	ZrO2		505	DLS	-23	Zeta-potential			-11.252	-3.19	-8.23	5.62	MTT
6843	21800953	ZrO2		505	DLS	-23	Zeta-potential			-11.252	-3.19	-8.23	5.62	MTT
6844														

- **Missing Data in Oxide NPs' Original Dataset (Dataset I)**
  - 18 % of Core Size Data
  - 39 % of Hydrodynamic Size Data
  - 41 % of Surface Charge Data
  - 74 % of Specific Surface Area Data



	No information of measurement	*Non-standard methods*	*Standard methods*
No information of data source	0	1	2
Reference	1	2	3
Manufacturer	2	3	4
Experiment	3	4	5

Reliability of data source

Reliability of measurement method

Quality & Completeness Assessment, Data Gap Filling and PChem score based Screening

Data Gap Filling Method – Nano Read-Across

Missing data replacement

Conventional Approach

- substitute missing values with mean values of non-missing values

Nano Read Across

- estimation from other properties of the same nanomaterials (e.g., estimating specific surface area from core size).

- information form manufacturer's specification sheet or other references using the same nanomaterials

SCIENTIFIC REPORTS

OPEN Toxicity Classification of Oxide Nanomaterials: Effects of Data Gap Filling and PChem Score-based Screening Approaches

Ha et al.(2018)
(Impact Factor = 4.259)

Environmental Science Nano

Trinh et al (2018) (IF = 6.047)

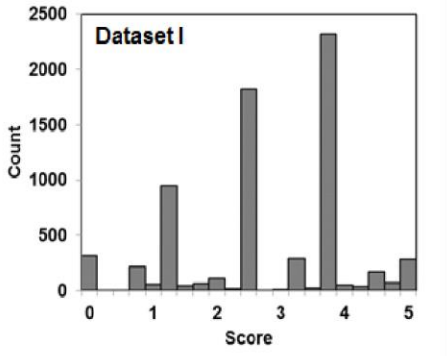
Curation of datasets, assessment of their quality and completeness, and nanoSAR classification model development for metallic nanoparticles†

Tung X. Trinh, [a] My Kieu Ha,[a] Jang Sik Choi,[b] Hyung Gi Byun [b] and Tae Hyun Yoon *[a]

Phem Data Quality Score based Screening

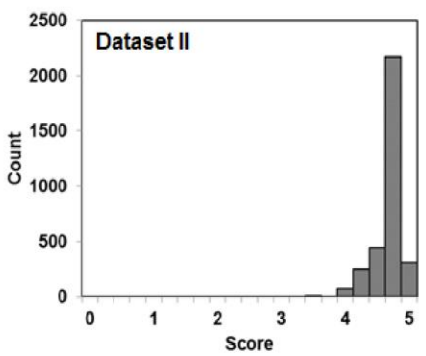
Dataset Ⅰ
- Orginal dataset
- Mean substitution

6842 rows
score = 2.8 ±1.3



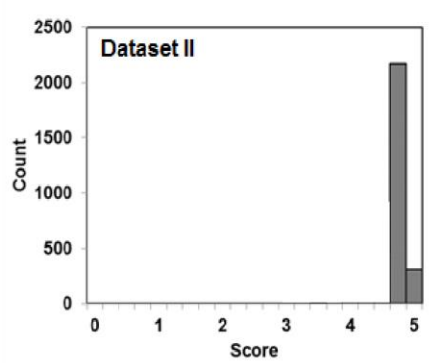
Dataset Ⅱ
- Data gap filling

3246 rows
score = 4.7 ± 0.2



DatasetⅢ
- Data gap filling
- 50% data with top PChem score

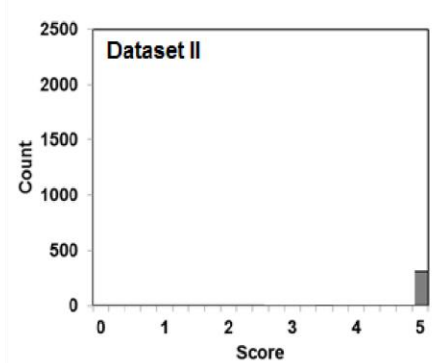
1738 rows
score = 4.8 ±0.1



Top 50%

DatasetⅢ
- Data gap filling
- 20% data with top PChem score

666 rows
score = 4.9 ±0.2



Top 20%

SCIENTIFIC REPORTS

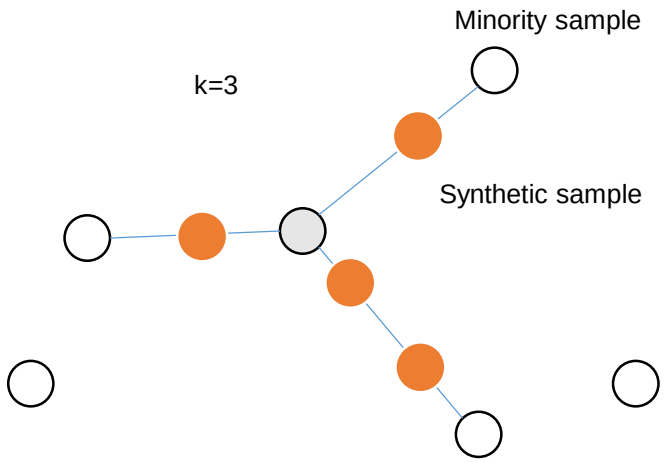
Towards a generalized toxicity prediction model for oxide nanomaterials using integrated data from different sources

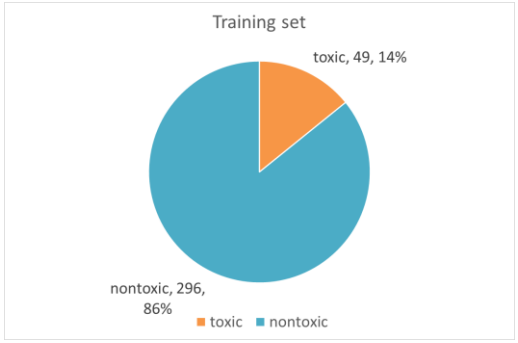
Jang-Sik Choi[1], My Kieu Ha[2], Tung Xuan Trinh[2], Tae Hyun Yoon[2] & Hyung-Gi Byun[1]

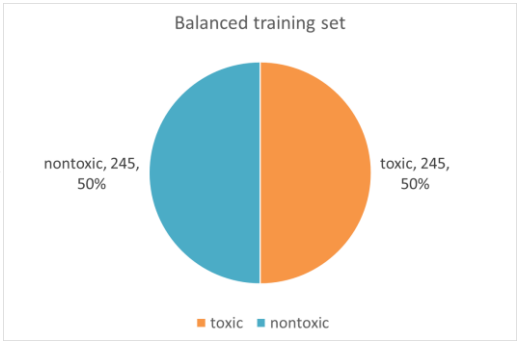
Choi et al (2018)

Data Imbalance Issue

- **SMOTE (Synthetic minority over-sampling technique)**







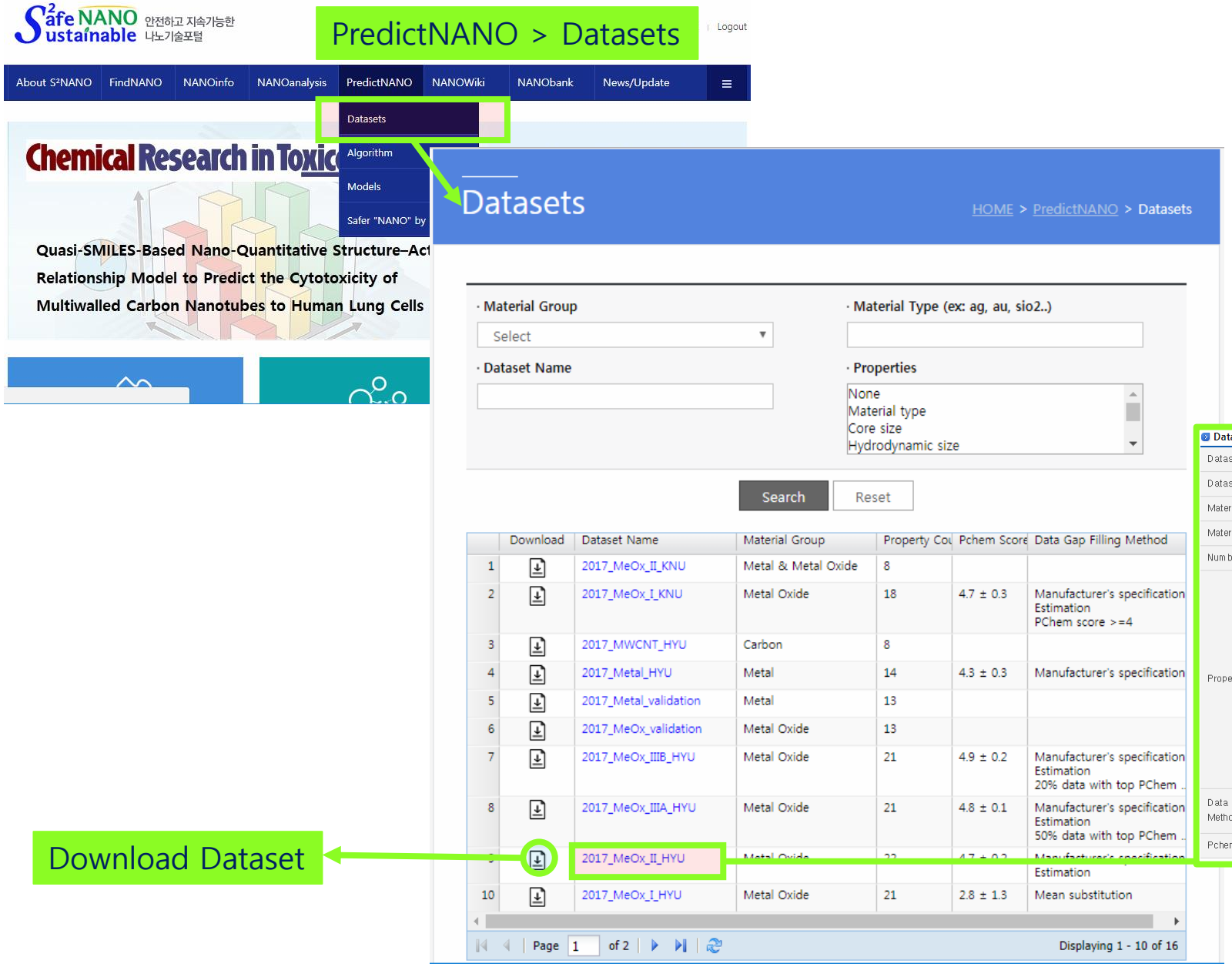
Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Toxic 16% : Nontoxic 84 %

SMOTE (Synthetic Minority Over-sampling TEchnique)

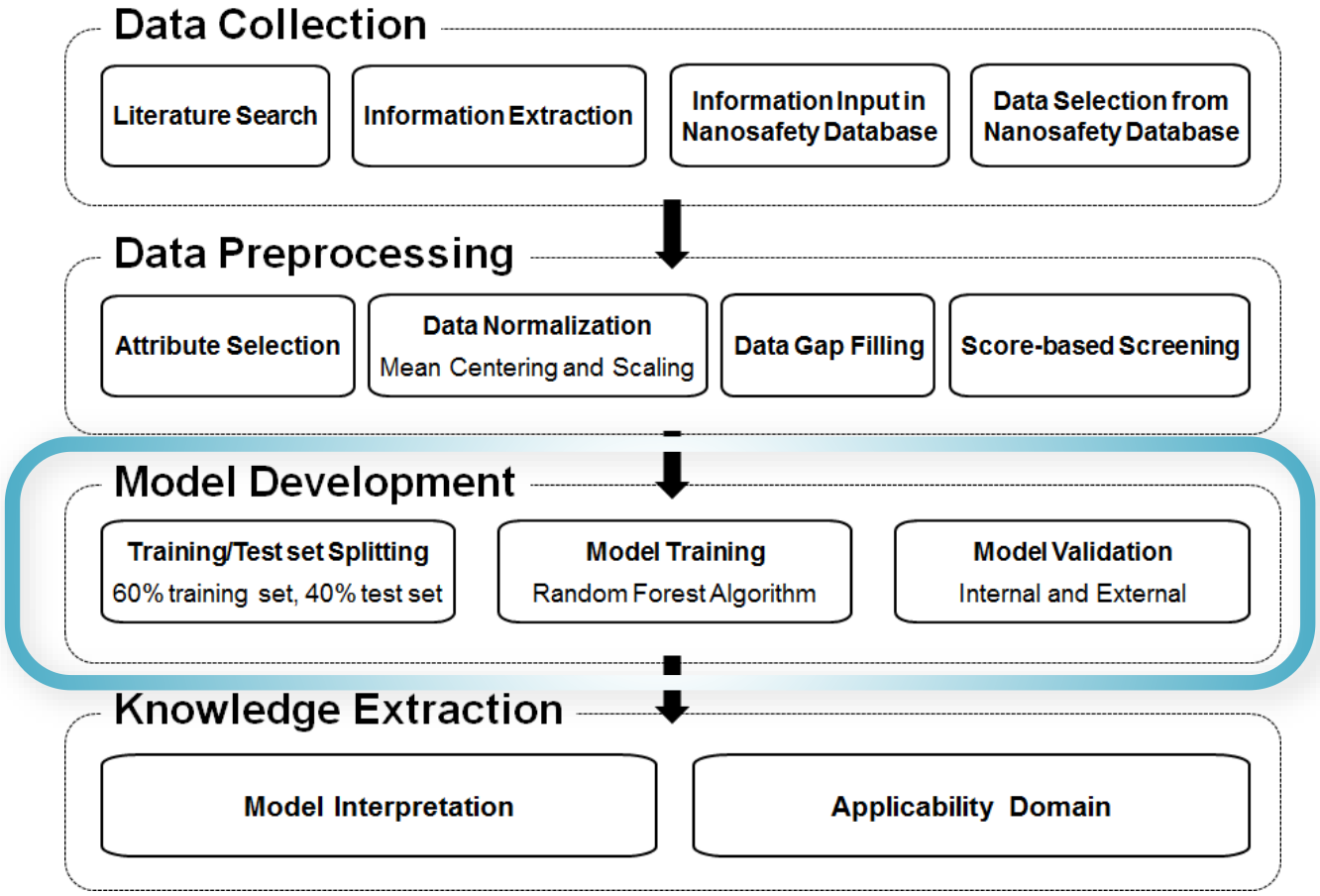
ID (Imbalanced Data) vs. BD (Balanced Data)

Curated Datasets

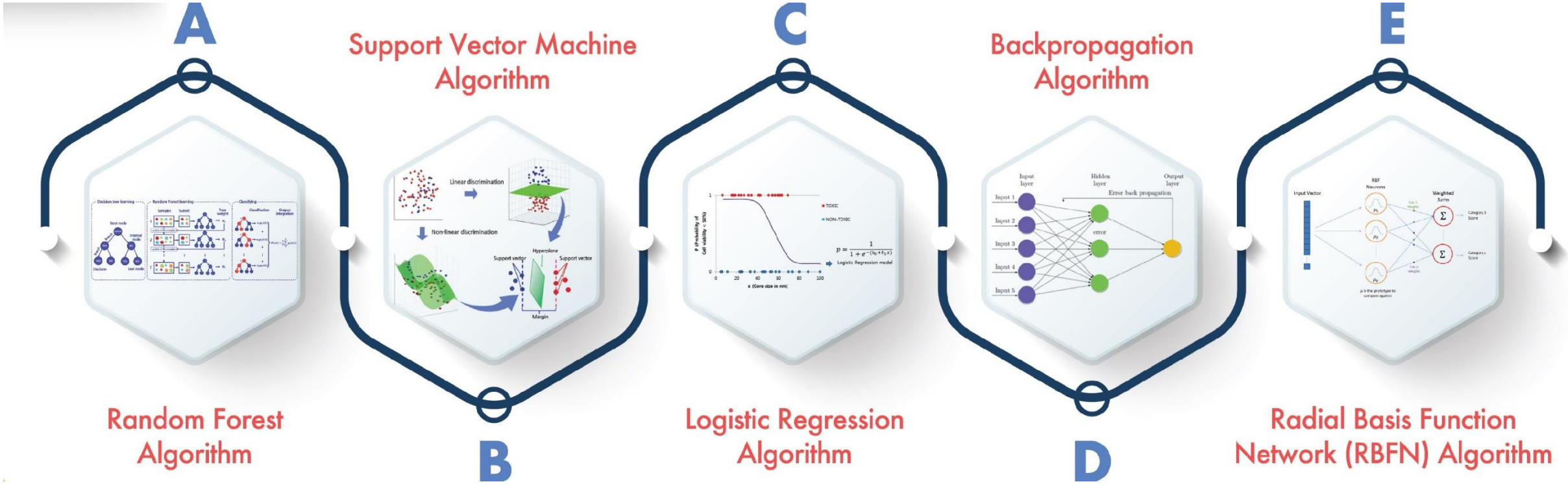


Model Development Workflow in S²NANO

Logistic Regression Algorithm
Random Forest Algorithm
Support Vector Machine Algorithm
Backpropagation Algorithm



Model Development – Algorithm Selection



Model Development – Validation (Internal & External)

Model validation

Validation	Dataset		Purpose
Internal validation (10 fold cross validation)	Training set (normalized by each method)	Balanced training set (by SMOTE)	1. Select the normalization method appropriate for each modeling algorithm 2. Look at SMOTE effect
External validation	Test set	-	1. Look at SMOTE effect 2. Select the best predictive model
Reliability validation	Cytotoxicity data measured by experiment	-	Validate the reliability of the best predictive model

Performance measures

		True condition	
		positive toxic	negative nontoxic
Predicted condition	positive toxic	True Positive (TP)	False Negative (FN)
	negative nontoxic	False Positive (FP)	True Negative (TN)
		$\text{Sensitivity} = \frac{TP}{TP + FP}$	$\text{Specificity} = \frac{TN}{FN + TN}$

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

Balanced accuracy

$$= \frac{1}{2}\left(\frac{TP}{TP + FP} + \frac{TN}{FN + TN}\right)$$

SCIENTIFIC REPORTS

Towards a generalized toxicity prediction model for oxide nanomaterials using integrated data from different sources

Jang-Sik Choi[1], My Kieu Ha[2], Tung Xuan Trinh[2], Tae Hyun Yoon[2] & Hyung-Gi Byun[1]

Choi et al (2018)

Model Development – Performance Measures

SCIENTIFIC REPORTS

Towards a generalized toxicity prediction model for oxide nanomaterials using integrated data from different sources

Jang-Sik Choi[1], My Kieu Ha[2], Tung Xuan Trinh[2], Tae Hyun Yoon[2] & Hyung-Gi Byun[1]

Choi et al (2018)

Normalization Method

Algorithm	Normalization method	True positive	False positive	False negative	True negative	Sensitivity	Specificity	Balanced accuracy	Standard deviation		
									Sensitivity	Specificity	Balanced accuracy
LR	min-max	39	12	16	278	70.91%	95.86%	83.39%	4.84%	1.18%	2.45%
	z-score	39	12	16	278	70.91%	95.86%	83.39%	4.84%	1.58%	2.47%
	log	46	11	9	279	83.64%	96.21%	89.92%	4.45%	1.01%	2.33%
	combination	45	8	10	282	81.82%	97.24%	89.53%	3.99%	1.15%	3.99%
SVM	min-max	28	6	27	284	50.91%	97.93%	74.42%	10.07%	0.72%	5.04%
	z-score	29	7	26	283	52.73%	97.59%	75.16%	9.73%	0.35%	4.87%
	log	40	5	15	285	72.73%	98.28%	85.50%	5.48%	0.65%	2.65%
	combination	41	5	14	285	74.55%	98.28%	86.41%	5.48%	0.67%	2.73%
RF	min-max	45	5	10	285	81.82%	98.28%	90.05%	5.71%	0.64%	2.92%
	z-score	44	5	11	285	80.00%	98.28%	89.14%	4.32%	0.67%	2.25%
	log	45	5	10	285	81.82%	98.28%	90.05%	5.02%	0.61%	2.49%
	combination	45	5	10	285	81.82%	98.28%	90.05%	4.93%	0.61%	2.44%
ANN	min-max	38	15	17	275	69.09%	94.83%	81.96%	15.31%	1.07%	7.79%
	z-score	40	6	15	284	72.73%	97.93%	85.33%	5.89%	0.60%	2.98%
	log	43	8	12	282	78.18%	97.24%	87.71%	6.00%	0.64%	3.01%
	combination	48	8	7	282	87.27%	97.24%	92.26%	4.05%	0.66%	2.11%

SCIENTIFIC REPORTS

Towards a generalized toxicity prediction model for oxide nanomaterials using integrated data from different sources

Jang-Sik Choi[1], My Kieu Ha[2], Tung Xuan Trinh[2], Tae Hyun Yoon[2] & Hyung-Gi Byun[1]

Choi et al (2018)

Data Imbalance Issue

Normalization method	Algorithm	Data	True positive	False positive	False negative	True negative	Sensitivity	Specificity	Balanced accuracy	Standard deviation		
										Sensitivity	Specificity	Balanced accuracy
Log	LR	ID	46	11	9	279	83.64%	96.21%	89.92%	4.45%	1.01%	2.33%
		BD	243	27	18	234	93.10%	89.66%	91.38%	5.84%	9.31%	7.25%
Combination	SVM	ID	41	5	14	285	74.55%	98.28%	86.41%	5.48%	0.67%	2.73%
		BD	255	7	6	254	97.70%	97.32%	97.51%	1.15%	0.98%	0.56%
Combination	RF	ID	45	5	10	285	81.82%	98.28%	90.05%	4.93%	0.61%	2.44%
		BD	253	4	8	257	96.93%	98.47%	97.70%	0.88%	0.72%	0.63%
Combination	ANN	ID	48	8	7	282	87.27%	97.24%	92.26%	4.05%	0.66%	2.11%
		BD	258	5	3	256	98.85%	98.08%	98.47%	0.79%	0.85%	0.75%

Toxic 16% : Nontoxic 84 %

SMOTE (Synthetic Minority Over-sampling TEchnique)

ID (Imbalanced Data) vs. BD (Balanced Data)

Towards a generalized toxicity prediction model for oxide nanomaterials using integrated data from different sources

Jang-Sik Choi[1], My Kieu Ha[2], Tung Xuan Trinh[2], Tae Hyun Yoon[2] & Hyung-Gi Byun[1]

SCIENTIFIC REPORTS

Choi et al (2018)

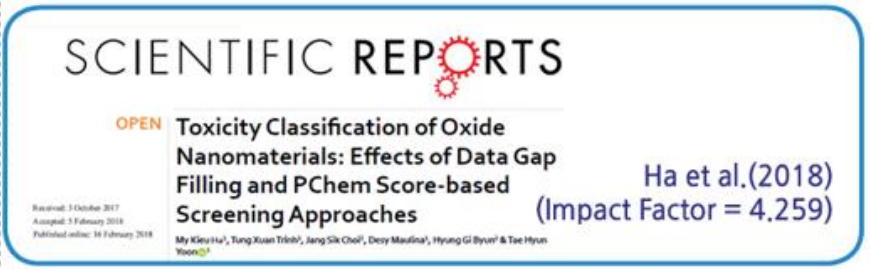
Internal vs. External Validations

Internal validation

Normalization method	Algorithm	Data	True positive	False positive	False negative	True negative	Sensitivity	Specificity	Balanced accuracy	Standard deviation		
										Sensitivity	Specificity	Balanced accuracy
Log	LR	ID	46	11	9	279	83.64%	96.21%	89.92%	4.45%	1.01%	2.33%
		BD	243	27	18	234	93.10%	89.66%	91.38%	5.84%	9.31%	7.25%
Combination	SVM	ID	41	5	14	285	74.55%	98.28%	86.41%	5.48%	0.67%	2.73%
		BD	255	7	6	254	97.70%	97.32%	97.51%	1.15%	0.98%	0.56%
Combination	RF	ID	45	5	10	285	81.82%	98.28%	90.05%	4.93%	0.61%	2.44%
		BD	253	4	8	257	96.93%	98.47%	97.70%	0.88%	0.72%	0.63%
Combination	ANN	ID	48	8	7	282	87.27%	97.24%	92.26%	4.05%	0.66%	2.11%
		BD	258	5	3	256	98.85%	98.08%	98.47%	0.79%	0.85%	0.75%

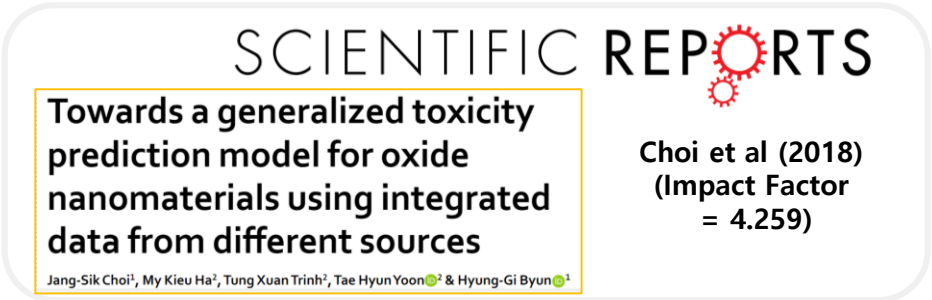
External validation

Normalization method	Algorithm	Data	True positive	False positive	False negative	True negative	Sensitivity	Specificity	Balanced accuracy	Standard deviation		
										Sensitivity	Specificity	Balanced accuracy
Log	LR	ID	25	6	4	194	86.21%	97.00%	91.60%	7.95%	0.70%	3.86%
		BD	26	21	3	179	89.66%	89.50%	89.58%	9.10%	7.49%	5.46%
Combination	SVM	ID	22	5	7	195	75.86%	97.50%	86.68%	11.64%	1.06%	5.71%
		BD	25	10	4	190	86.21%	95.00%	90.60%	8.85%	1.85%	3.79%
Combination	RF	ID	24	3	5	197	82.76%	98.50%	90.63%	12.75%	1.12%	6.23%
		BD	25	9	4	191	86.21%	95.50%	90.85%	11.24%	1.84%	5.84%
Combination	ANN	ID	23	4	6	196	79.31%	98.00%	88.66%	8.74%	1.42%	4.42%
		BD	27	13	2	187	93.10%	93.50%	93.30%	6.41%	0.83%	3.08%



Performance Comparisons of nanoSAR classification Models

Model Name	Dataset	Algorithm	Internal validation	External validation	Publications
			Accuracy or R^2	Accuracy or R^2	
PM-100	2017_Metal_HYU	Random forest	86%	82%	
PM-101	2016_Metal_KNU	Backpropagation	95%	67%	
PM-102	2015_Metal_HYU	Support vector machine	90%	75%	
PM-103	2015_Metal_MeOx_KNU	Backpropagation	91%	86%	
PM-104	2015_Metal_HYU	Support vector machine	90%	82%	
PM-106	2017_MeOx_I_KNU	Resilient backpropagation	93%	86%	
PM-107	2017_MeOx_II_HYU	Random forest	94%	54%	Ha et al. (2018) Scientific Reports
PM-108	2016_MeOx_KNU	Backpropagation	95%	85%	Choi et al. (2018) Scientific Reports
PM-114	2017_MeOx_I_KNU	Random forest	91%	87%	
PM-115	2017_MeOx_I_KNU	Support vector machine	93%	89%	
PM-116	2017_MWCNT_HYU	Quasi-QSAR	0.89	N/A	Trinh et al. (2018) Chem. Res. In Toxicology
PM-117	2017_MeOx_II_KNU	Quasi-QSAR	0.79	0.71	Choi et al. (2018) Chemosphere
PM-118	2015_Metal_KNU	Backpropagation	88%	67%	
PM-120	2017_Metal_HYU	Support vector machine	86%	82%	Trinh et al. (2018) Environmental Science : NANO

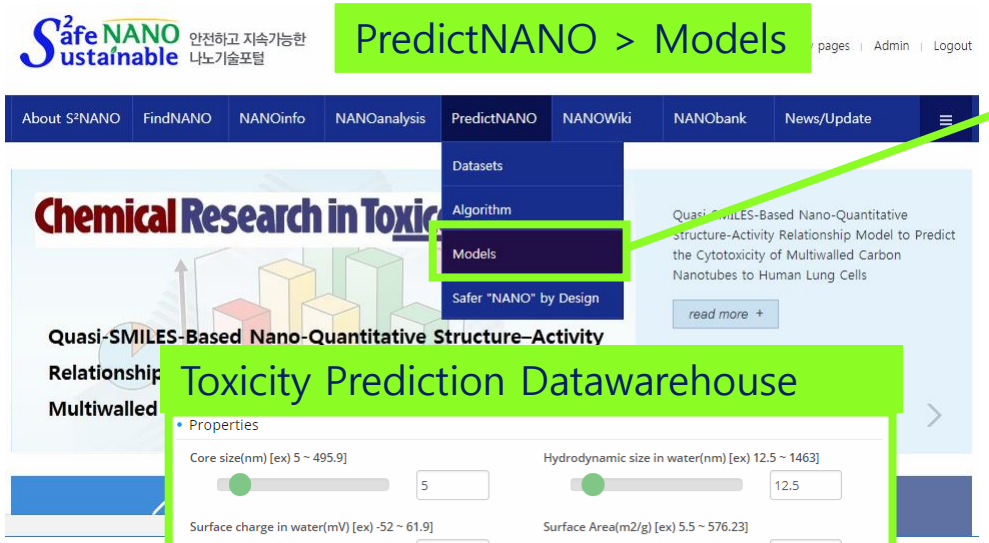


Choi et al (2018) (Impact Factor = 4.259)

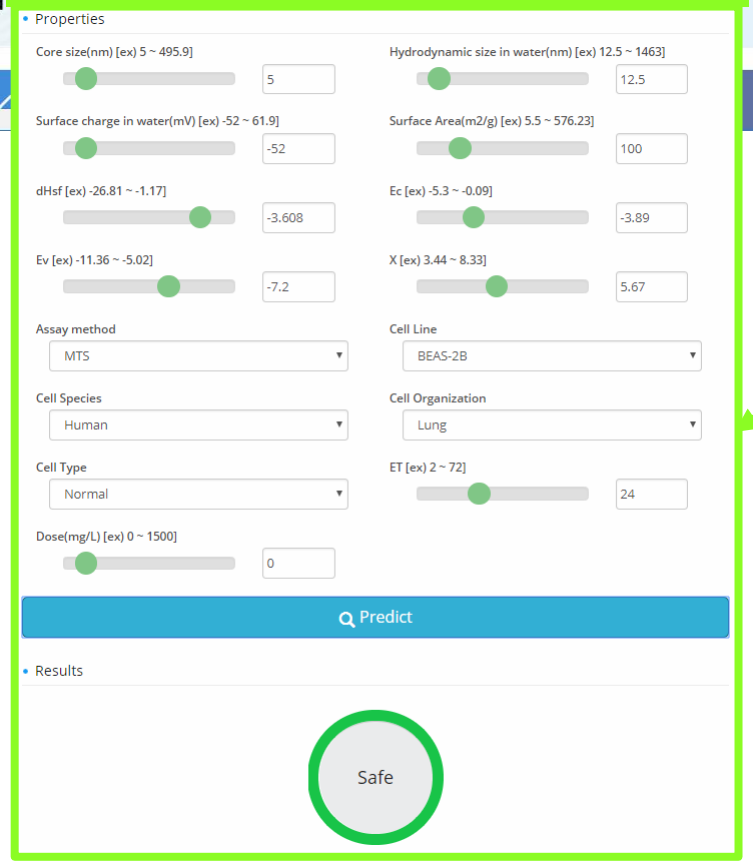
Trinh et al (2018) (Impact Factor = 6.047)

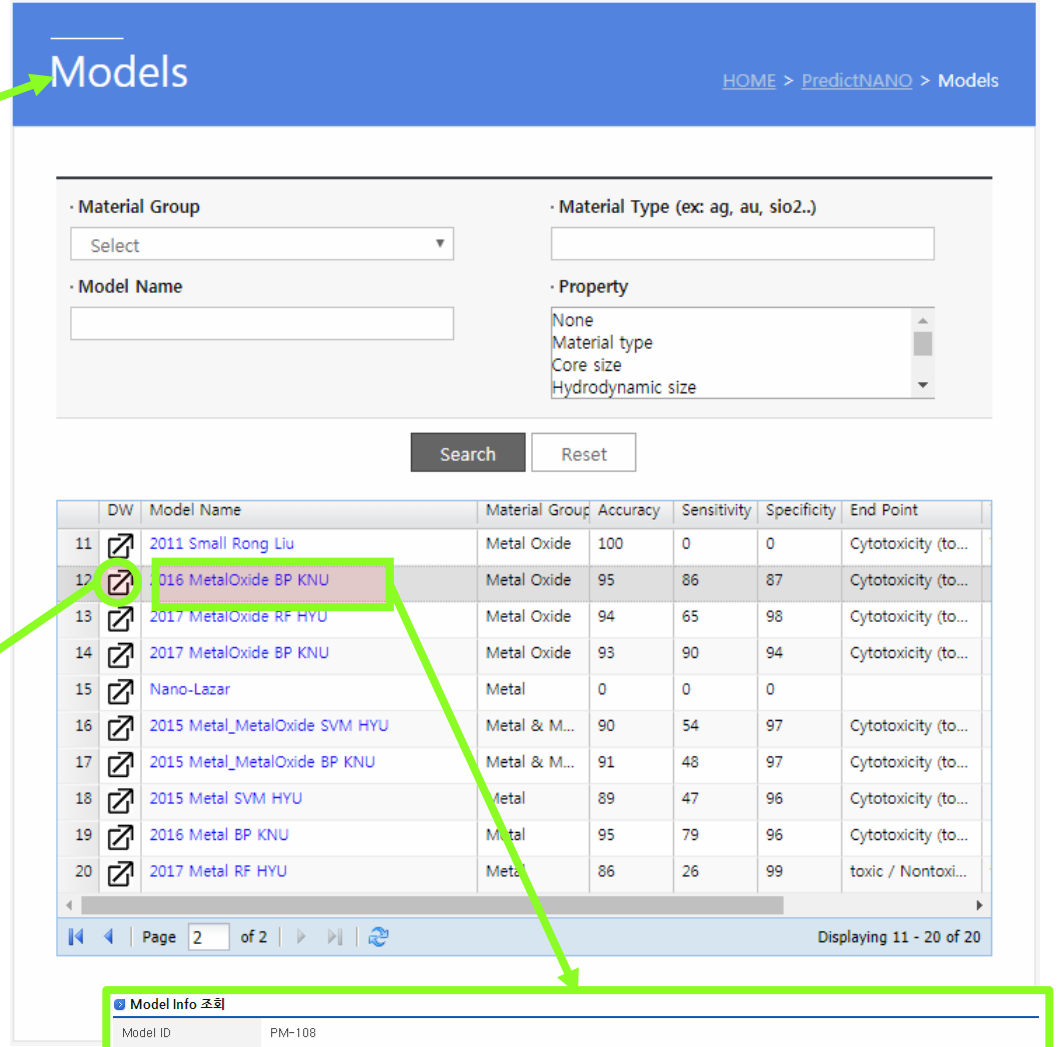
Prediction Models

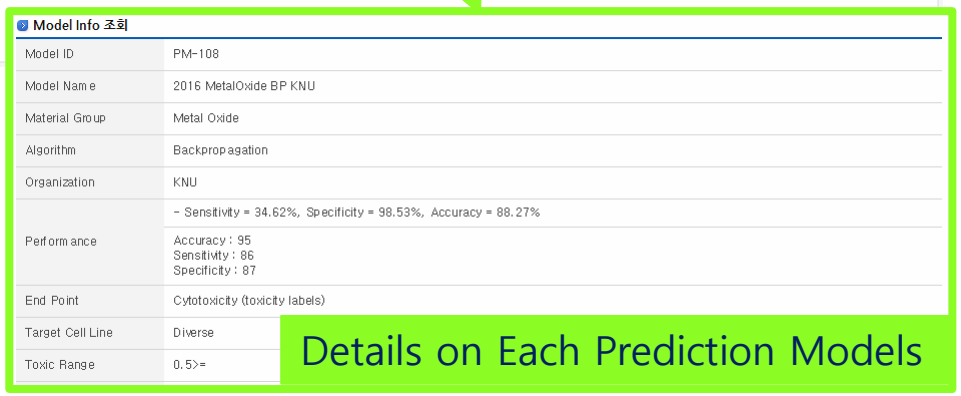
PredictNANO > Models



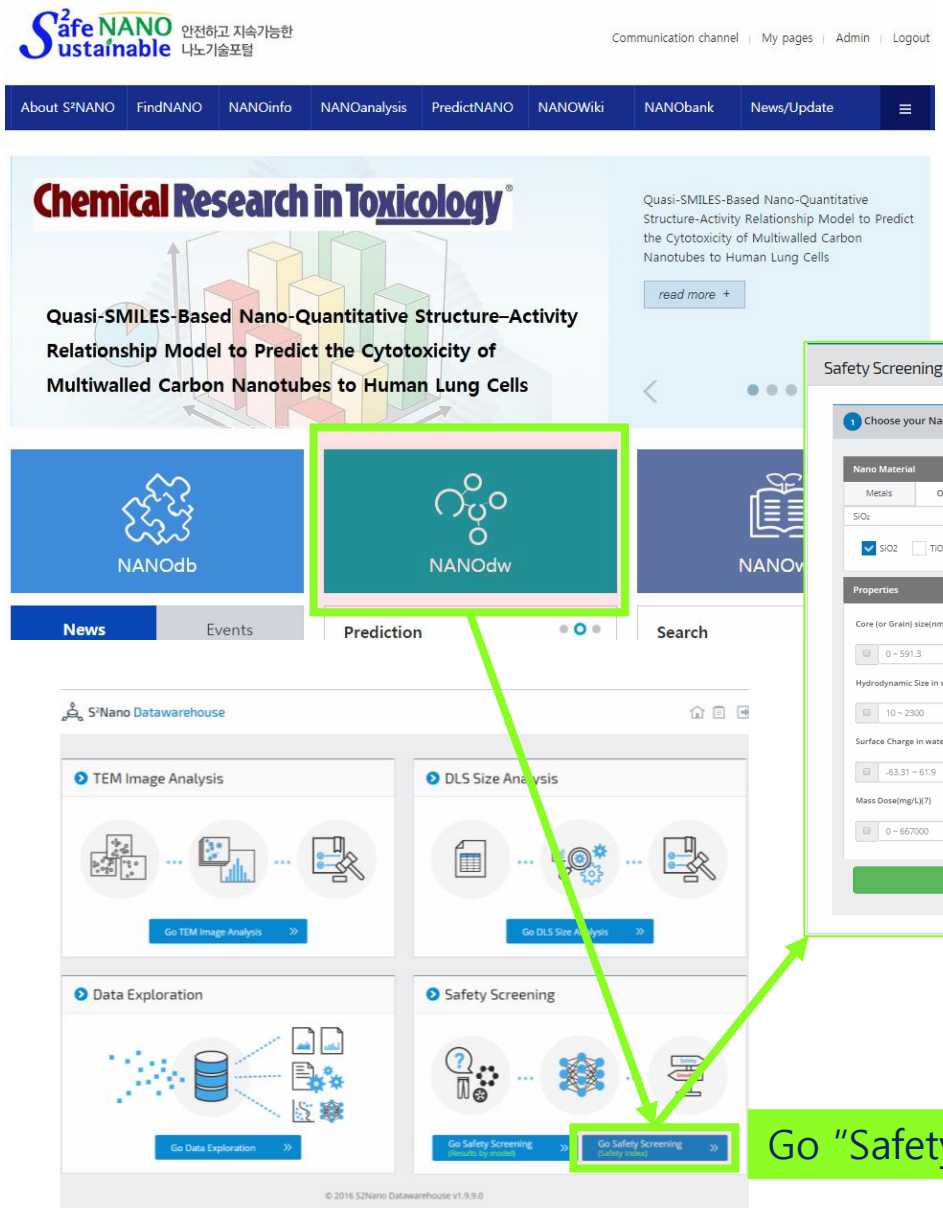
Toxicity Prediction Datawarehouse

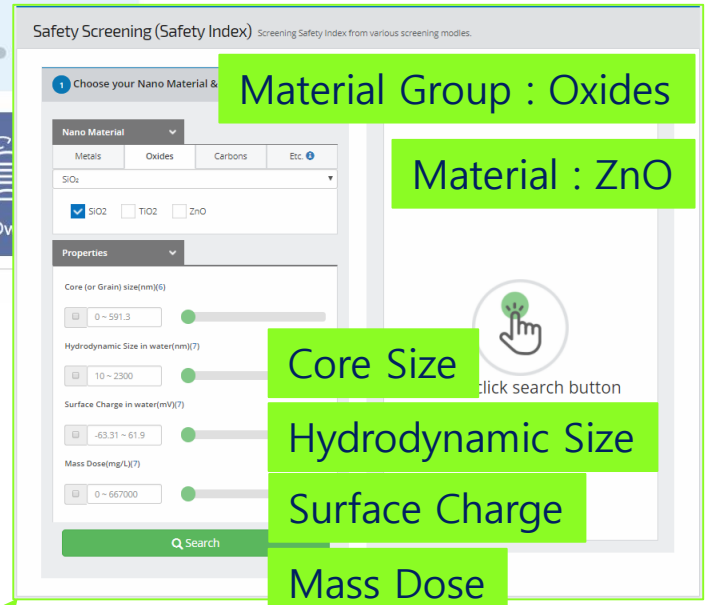






Details on Each Prediction Models





Material Group : Oxides

Material : ZnO

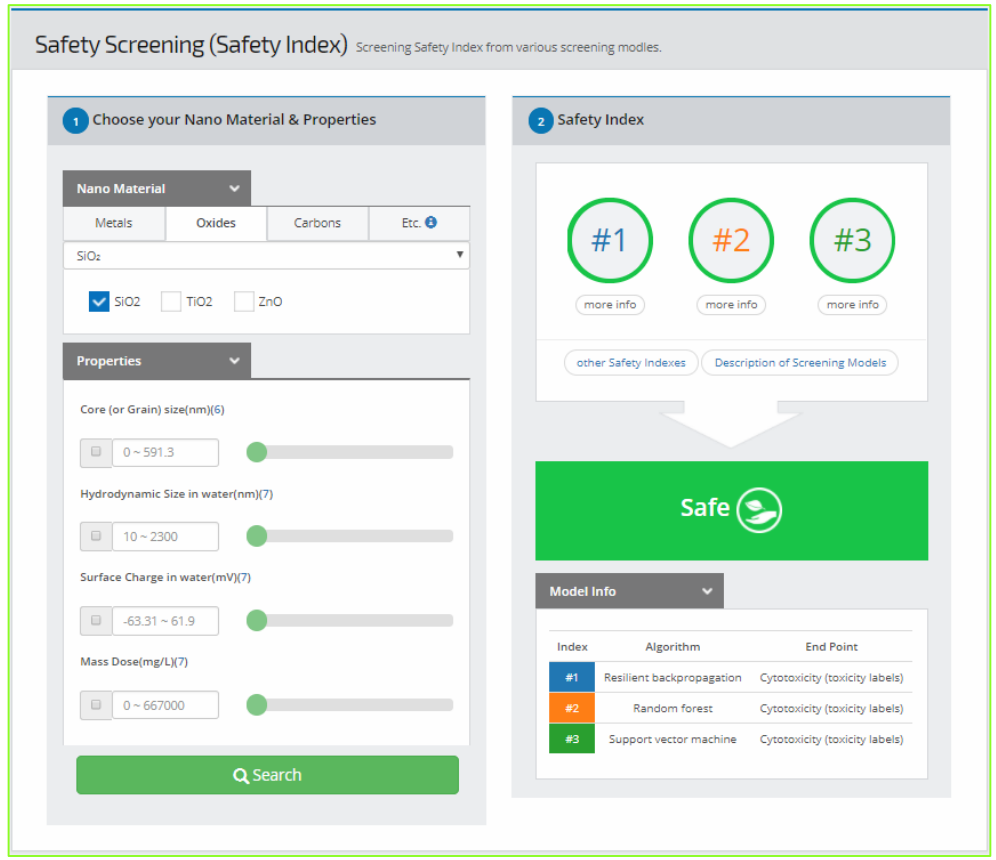
Core Size

Hydrodynamic Size

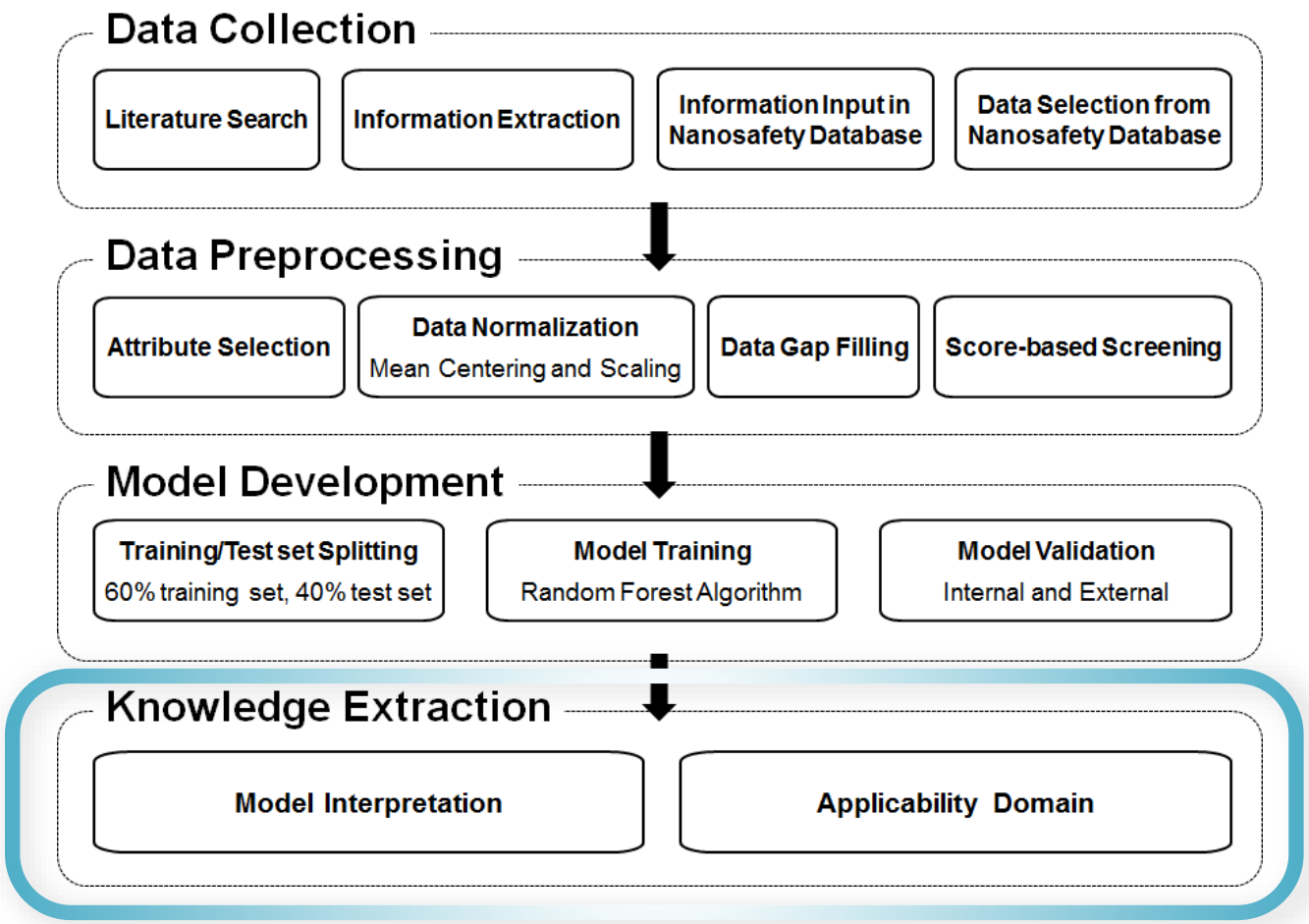
Surface Charge

Mass Dose

Go "Safety Screening"



Model Development Workflow in S²NANO



Relative Importance of Attributes

Applicability Domains

Attribute Importance

SCIENTIFIC REPORTS

OPEN Toxicity Classification of Oxide Nanomaterials: Effects of Data Gap Filling and PChem Score-based Screening Approaches

Ha et al.(2018) (Impact Factor = 4.259)

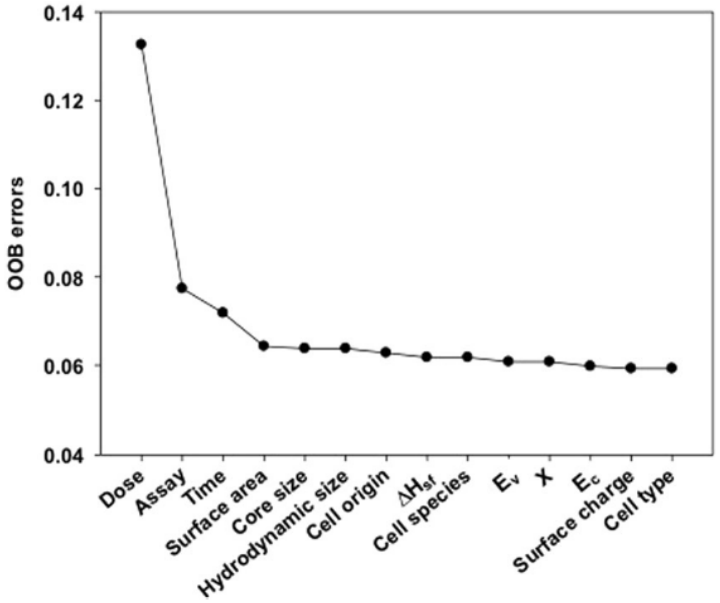


Figure 4. Leave-one-out OOB errors against attributes.

Environmental Science Nano

Trinh et al (2018) (Impact Factor = 6.047)

Curation of datasets, assessment of their quality and completeness, and nanoSAR classification model development for metallic nanoparticles†

Tung X. Trinh, [a] My Kieu Ha,[a] Jang Sik Choi,[b] Hyung Gi Byun [b] and Tae Hyun Yoon *[a]

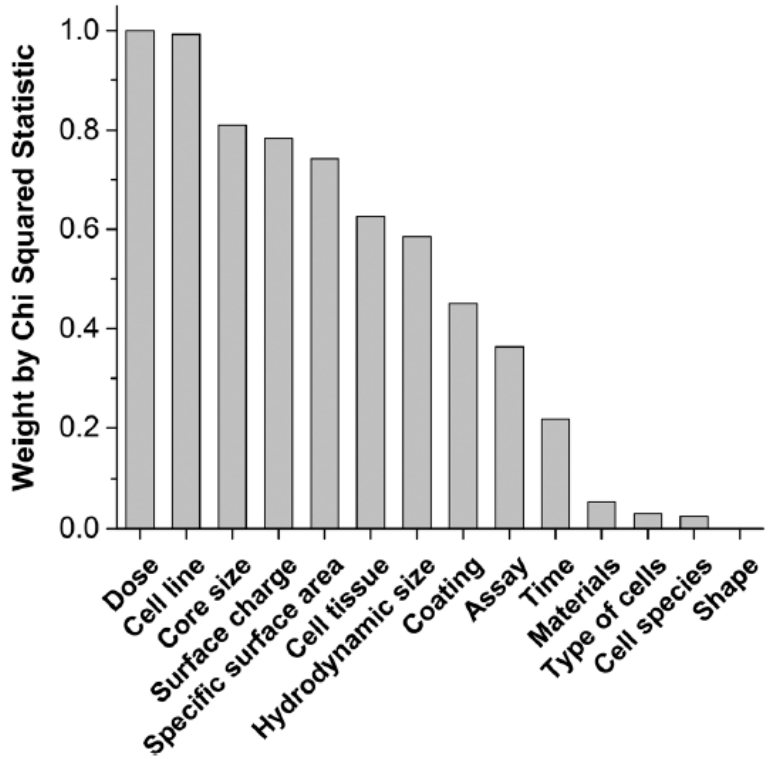


Fig. 3 Attribute importance to the toxicity of metallic NPs. Importance of each attribute to toxicity is represented by their weights

Attribute	Relative importance
Dose	11.10
ΔHsf	7.34
Exposure time	5.33
Hydrodynamic size	4.43
Ec	3.66
Surface area	3.58
Core size	3.57
Cell species	2.88
xMeO	2.44
Cell type	2.44
Surface charge	1.90
Assay method	1.82
Ev	1.45
Cell name	1.33
Cell origin	1.06

Table 8. Relative importance.

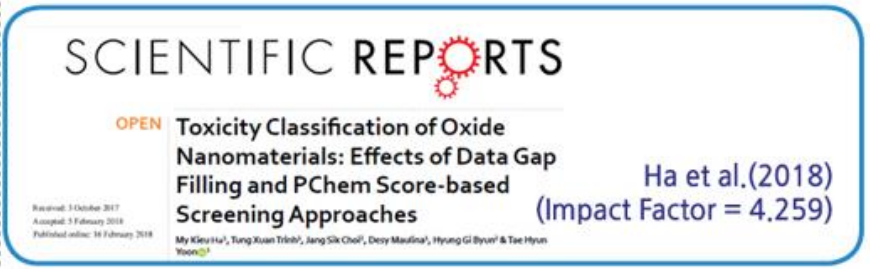
SCIENTIFIC REPORTS

Towards a generalized toxicity prediction model for oxide nanomaterials using integrated data from different sources

Jang-Sik Choi[1], My Kieu Ha[2], Tung Xuan Trinh[2], Tae Hyun Yoon[2] & Hyung-Gi Byun[1]

Choi et al (2018)

Applicability Domains of Models


SCIENTIFIC REPORTS

OPEN Toxicity Classification of Oxide Nanomaterials: Effects of Data Gap Filling and PChem Score-based Screening Approaches

Ha et al.(2018)
(Impact Factor = 4.259)

Attribute	I		II		III-A		III-B	
	Min.	Max.	Min.	Max.	Min.	Max.	Min.	Max.
Dose (μg/mL)	0	10000	0	167000	0	1500	0	1500
Time (h)	0	360	1	168	2	72	6	72
Core size (nm)	2.7	629	2.7	496	5	496	5.9	193
Hydro. size (nm)	8.6	6181	8.6	2300	12.5	1463	12.5	1457
Surface charge (mV)	−63.3	61.9	−63.3	61.9	−52	61.9	−47.6	42.8
Surface area (m^2/g)	0.8	1150	5.5	576	5.5	576	6	576
ΔH_{sf} (eV)	−64.7	−1.2	−64.7	−1.2	−26.8	−1.2	−26.8	−1.6
E_c (eV)	−6.6	−0.1	−6.6	−0.1	−5.2	−0.1	−5.3	−0.3
E_v (eV)	−11.4	−5.0	−11.3	−5.0	−11.1	−5.0	−11.4	−5.0
χ (eV)	3.2	8.3	3.4	8.3	3.4	6.8	3.8	8.3

Table 4. Applicability domains regarding the numerical attributes.


Environmental Science Nano

Trinh et al (2018)

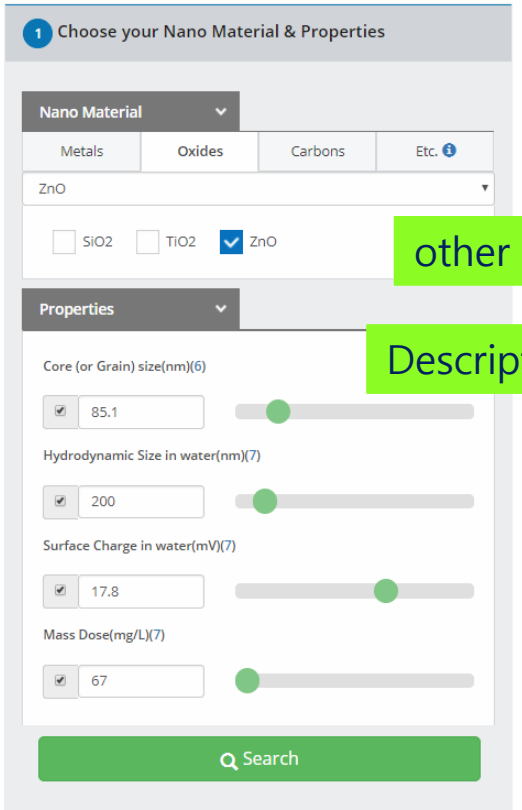
Curation of datasets, assessment of their quality and completeness, and nanoSAR classification model development for metallic nanoparticles†

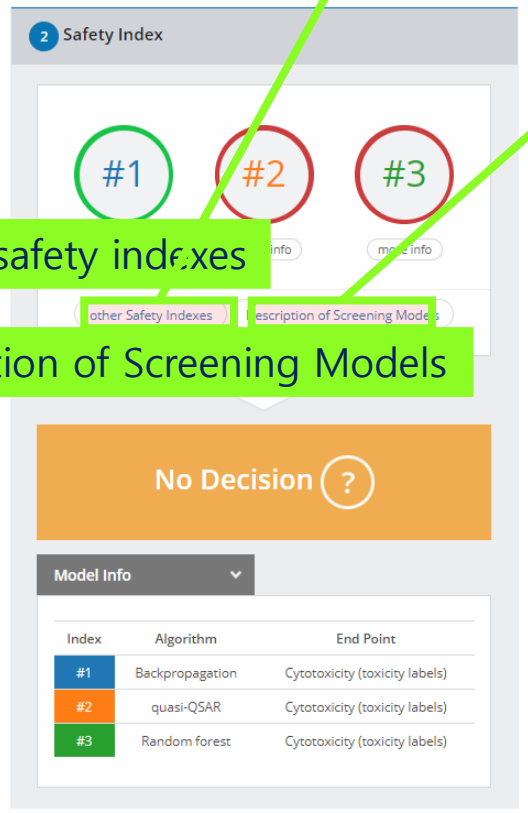
Tung X. Trinh, [a] My Kieu Ha,[a] Jang Sik Choi,[b] Hyung Gi Byun [b] and Tae Hyun Yoon *[a]

Table S6. Applicability domain of nanoSAR models built from datasets A, B and C.

Attribute	Dataset A	Dataset B	Dataset C
NPs type	Ag, Au	Ag, Au	Ag, Au
Shape	Particle, hollow, nanorod	Particle, hollow, nanorod	Particle, hollow, nanorod
Core size (nm)	2 – 120	2 – 120	2.5 - 120
Hydrodynamic size (nm)	7.1 – 300	7.1 – 300	7.1 – 300
Surface charge (mV)	-78.8 – 58.2	-78.8 – 58.2	-78.8 – 58.2
Specific surface area (m^2/g)	2.3 – 185.7	2.3 – 185.7	2.3
Dose (ppm)	0 – 400	0 – 400	0 – 400
Exposure time (h)	0 – 96	0 – 96	0 – 96

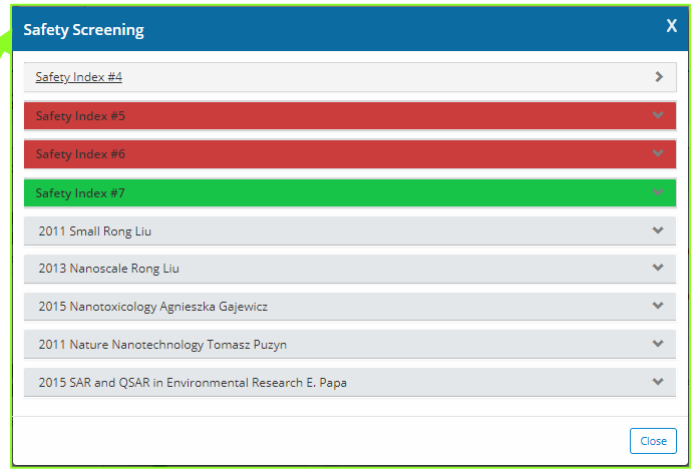
Datawarehouse with User-Friendly Interfaces

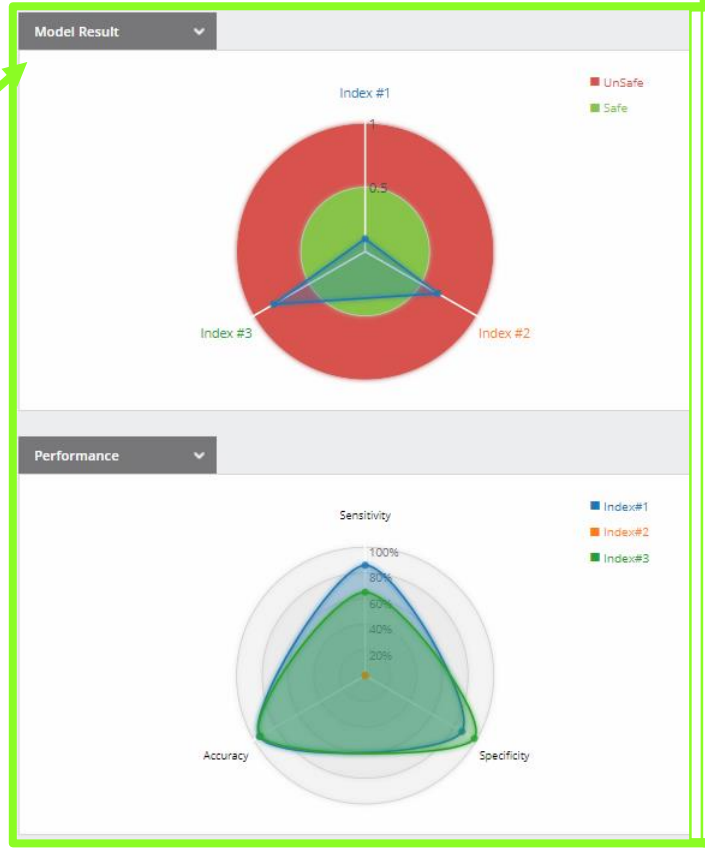




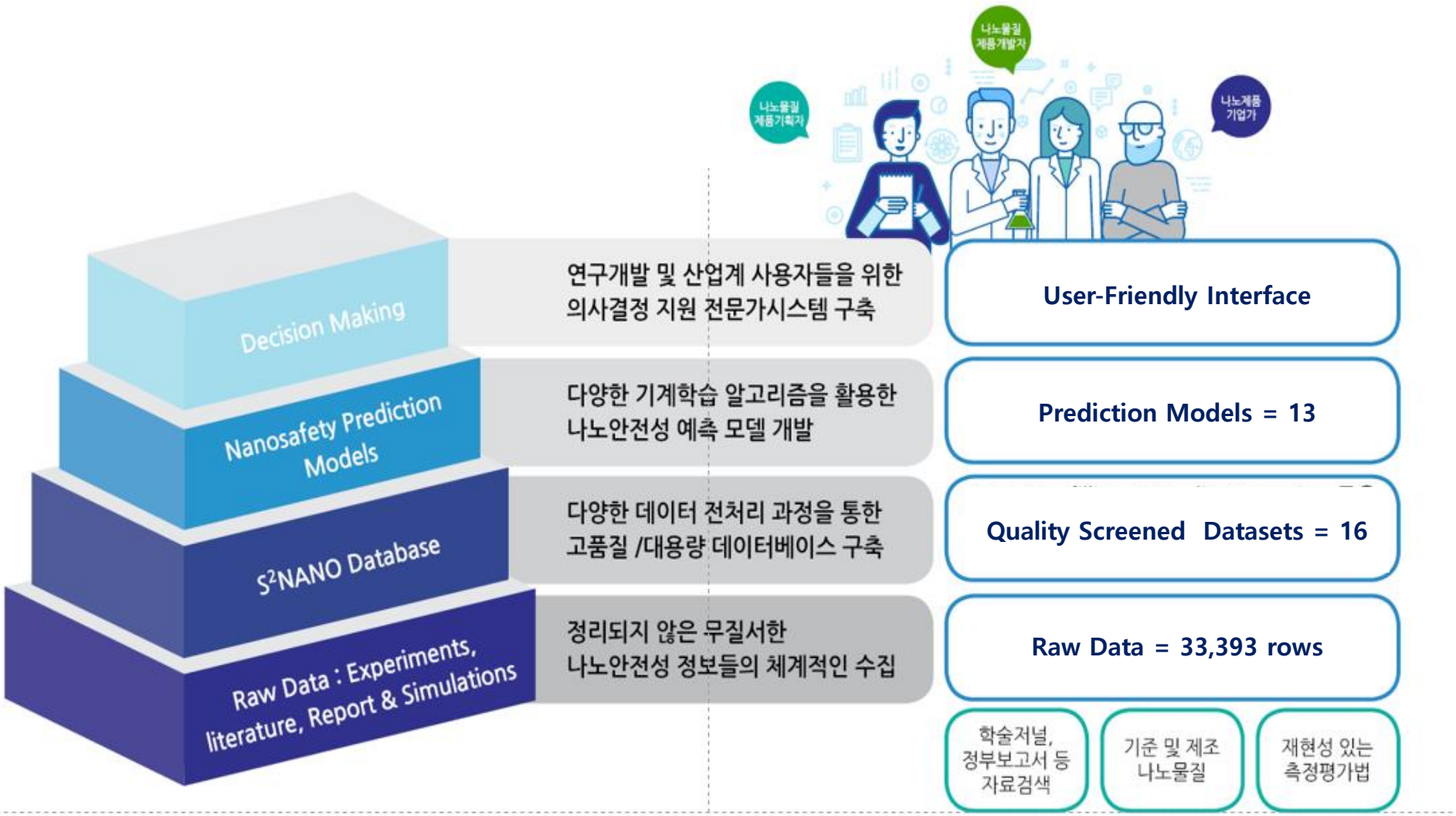
other safety indexes

Description of Screening Models





Implementation of Collected Database, Curated Datasets and nanoSAR classification models in S²NANO portal.

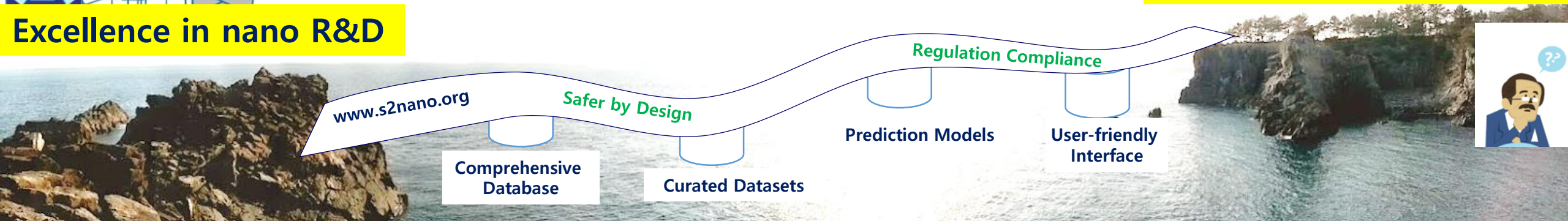


SUMMARY

- To overcome current issues in nanosafety data, such as small, unbalanced, & heterogeneous datasets with many missing values, we have collected **a comprehensive nanosafety database** (S2NANO) from experiments as well as literature mining. (33,393 rows of raw data were collected)
- These data were further processed and **16 quality screened datasets** were curated : Data gap-filled with nano read-across methods and assessed their data quality / completeness based on Pchem score. Using these curated datasets, **13 prediction models** were developed with different algorithms (LR, SVM, RF, ANN) and validated internally & externally.
- These comprehensive database, curated datasets, and nanosafety prediction models were **implemented in S2NANO portal with user-friendly interfaces** for future applications in safety by designand regulation compliance.

Excellence in nano R&D

Profitable nano Industry



ACEnano
Analytical and Characterisation Excellence

Measurement Techniques

Safe NANO Sustainable

http://portal.s2nano.org

Measurements & Models for Nanomaterials

NanoSolveIT

Predictive Models

QSAR vs Quasi-QSAR

[Problem] In the case of nanomaterials, the molecular structures of nanomaterials is the same as bulk chemicals

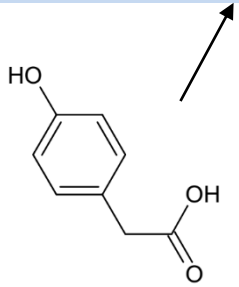
In the case of "classic" QSPR/QSAR analysis the paradigm is the following:
→ Endpoint is a mathematical function of molecular structure

Endpoint = Mathematical function (Molecular structure)

QSAR

SMILES

S_k
zone 1 zone 2 zone 3
c. . . . .
1. . . . .
(. . . . .
C. . . . .



Quasi-QSAR

[Solution] replace the traditional paradigm with by fresh paradigm

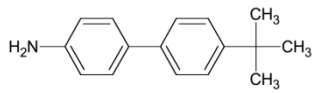
Endpoint = Mathematical function (Eclectic data)

Physicochemical characteristics

Biochemical conditions

Various conditions and characteristics of nanomaterial could impact associated biochemical endpoints!

Quasi-SMILES

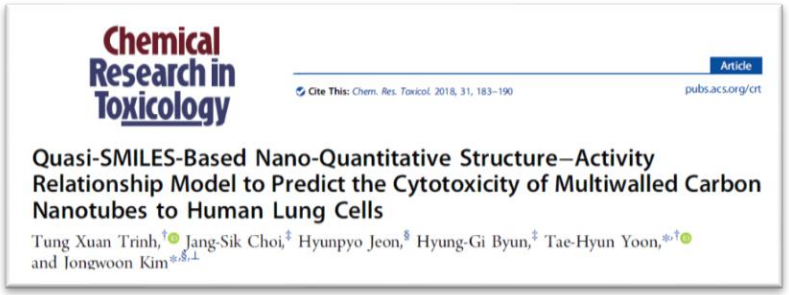


'T'

Dose		Code for Enginered size
200 μg/mL	"A"	A3
150 μg/mL	"B"	A3
100 μg/mL	"C"	A3
50 μg/mL	"D"	A3
		A3
		A3

Ref. Toropova, Alla P., and Andrey A. Toropov. "Optimal descriptor as a translator of eclectic information into the prediction of membrane damage by means of various TiO 2 nanoparticles." *Chemosphere* 93.10 (2013): 2650-2655.
Ref. Toropov, Andrey A., Robert Rallo, and Alla P. Toropova. "Use of quasi-SMILES and Monte Carlo optimization to develop quantitative feature property/activity relationships (QFPR/QFAR) for nanomaterials." *Current topics in medicinal chemistry* 15.18 (2015): 1837-1844.

Model Development – Performance Measures


Chemical Research in Toxicology

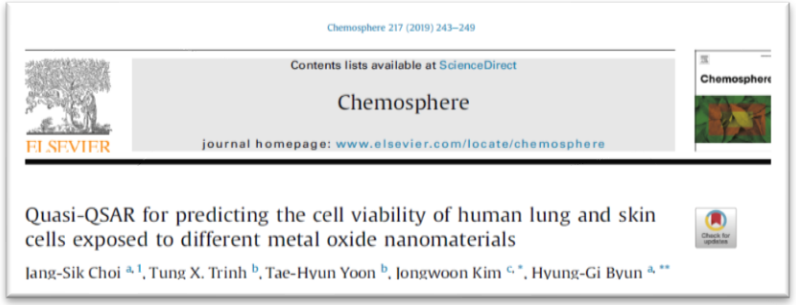
Article

Cite This: Chem. Res. Toxicol. 2018, 31, 183−190

pubs.acs.org/crt

Quasi-SMILES-Based Nano-Quantitative Structure−Activity Relationship Model to Predict the Cytotoxicity of Multiwalled Carbon Nanotubes to Human Lung Cells

Tung Xuan Trinh,[†] Jang-Sik Choi,[‡] Hyunpyo Jeon,[§] Hyung-Gi Byun,[‡] Tae-Hyun Yoon,[*,†] and Jongwoon Kim[*,§,⊥]


Chemosphere 217 (2019) 243–249

Contents lists available at ScienceDirect

Chemosphere

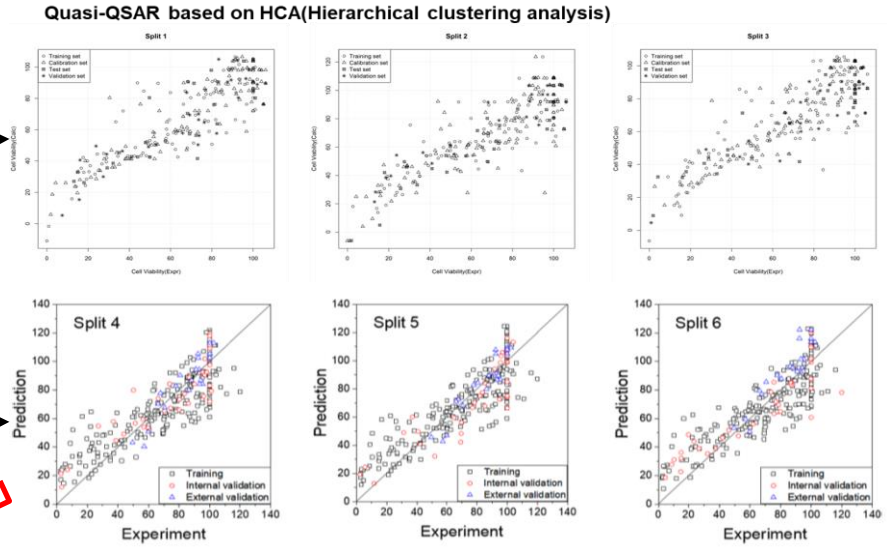
journal homepage: www.elsevier.com/locate/chemosphere

Quasi-QSAR for predicting the cell viability of human lung and skin cells exposed to different metal oxide nanomaterials

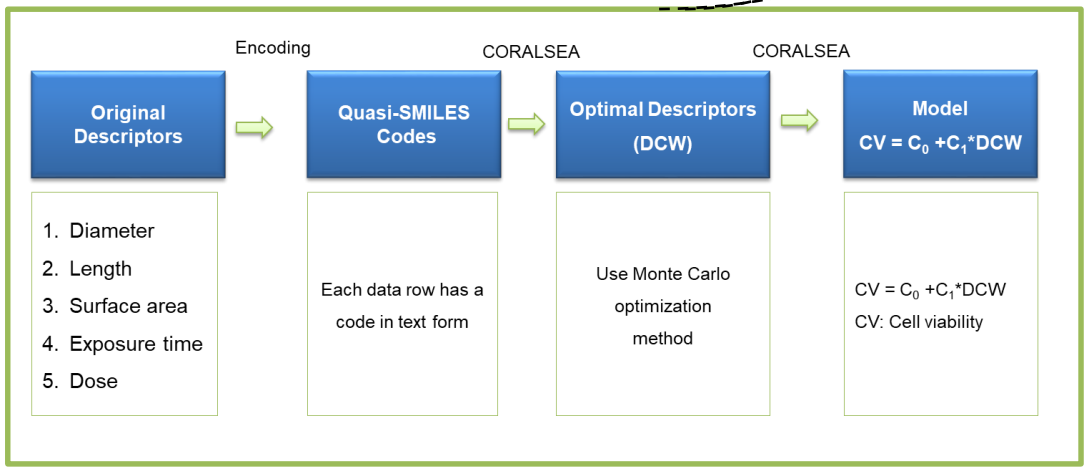
Jang-Sik Choi [a, 1], Tung X. Trinh [b], Tae-Hyun Yoon [b], Jongwoon Kim [c, *], Hyung-Gi Byun [a, **]

For MOs

For MWCNTs



Quasi-QSAR Modelling



For 20 MWCNTs, both HCA and Normalization showed good prediction results and HCA (R2: 0.83-0.91) outperformed Normalization (R2 0.70-0.78)

For 21 MOs, HCA (R2: 0.76-0.81) highly outperformed Normalization (R2: 046.-0.50)

Nano-QSARs based on quasi-SMILES were successfully developed for different MWCNTs and MOs by using HCA

The studies showed a potential of quasi-SMILES employing HCA overcomes the limitation in developing Nano-QSARs (i.e. increasing applicability domains of models)

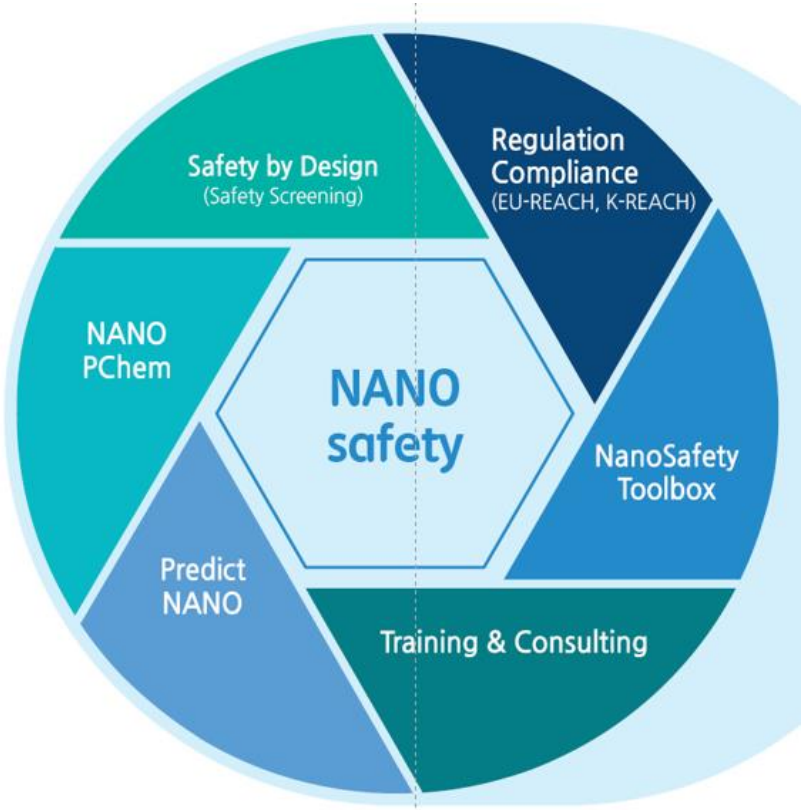
Acknowledgements

Safe NANO Sustainable
http://portal.s2nano.org

Ministry of Trade, Industry and Energy

ACEnano
Analytical and Characterisation Excellence

NanoSolveIT



TO TO21 Co.,Ltd. 한양대학교
나노안전성 DB 및
사용자 친화적 포털 고도화

KIT 안전성평가연구소 한양대학교
물성 및 독성 데이터 생산
고품질/대용량 나노안전성 데이터 생산

KNU 강원대학교 KIST Europe 한양대학교
나노안전성
예측 모델의 고도화

한양대학교 TO TO21 Co.,Ltd. KIST Europe
나노안전성 예측 시스템
최종 검증 및 향후
상용화 전략 수립

Acknowledgements

Collaborators

Prof. BYUN, Hyung-Gi (KNU)
Dr. CHOI, Jang-Sik (KNU/HYU)
Mr. CHOI, Woosu (TO21)
Mr. KIM, Doyoung (TO21)
Dr. BAE, Hee-Kyung (TO21)
Dr. KIM, Jongwoon (KIST-Europe/KRICT)
Dr. JEON, Hyun Pyo (KIST-Europe)
Dr. YOON, Seokju (KIT)
Dr. Oh, Jeong-Hwa (KIT)
Ms. HA, Kieu My (HYU)
Mr. TRINH, Xuan Tung (HYU)

Safe NANO Sustainable
http://portal.s2nano.org

TO21 Co.,Ltd. 한양대학교
나노안전성 DB 및
사용자 친화적 포털 고도화

KIT 안전성평가연구소 한양대학교
물성 및 독성 데이터 생산
고품질/대용량 나노안전성 데이터 생산

KNU 강원대학교 KIST Europe 한양대학교
나노안전성
예측 모델의 고도화

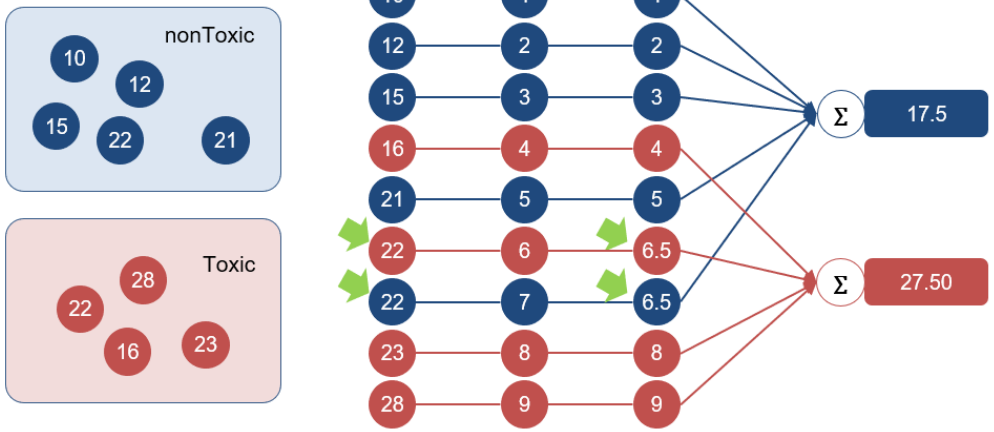
한양대학교 TO21 Co.,Ltd. KIST Europe
나노안전성 예측 시스템
최종 검증 및 향후
상용화 전략 수립

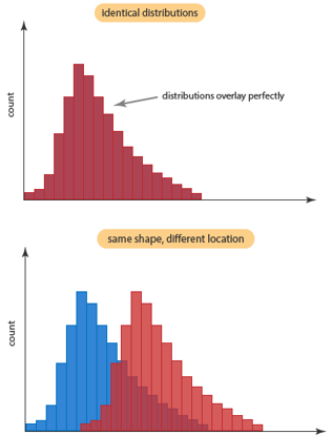
Thanks for your attention

Model Interpretation : Attribute Importance

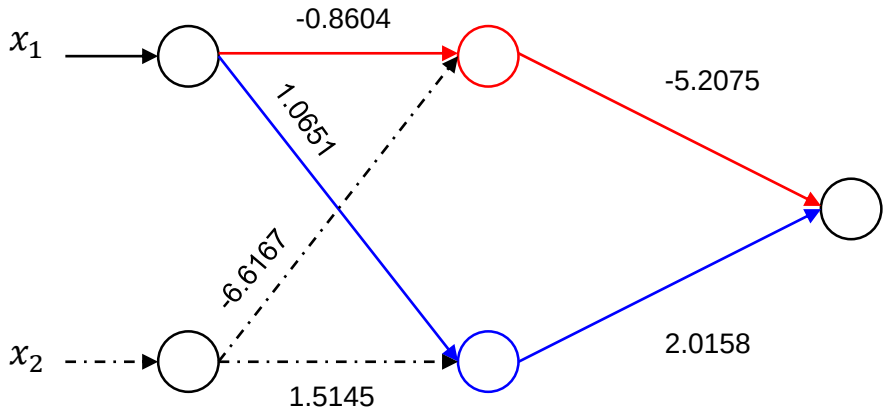
- **Mann-Whitney-Wilcoxon-test (also called the Wilcoxon rank-sum test)**

H0 : the distributions are the same
H1 : the distributions are not the same





- **Relative importance**



$$RI_x = \sum_{y=1}^{m} w_{xy} w_{yz}$$

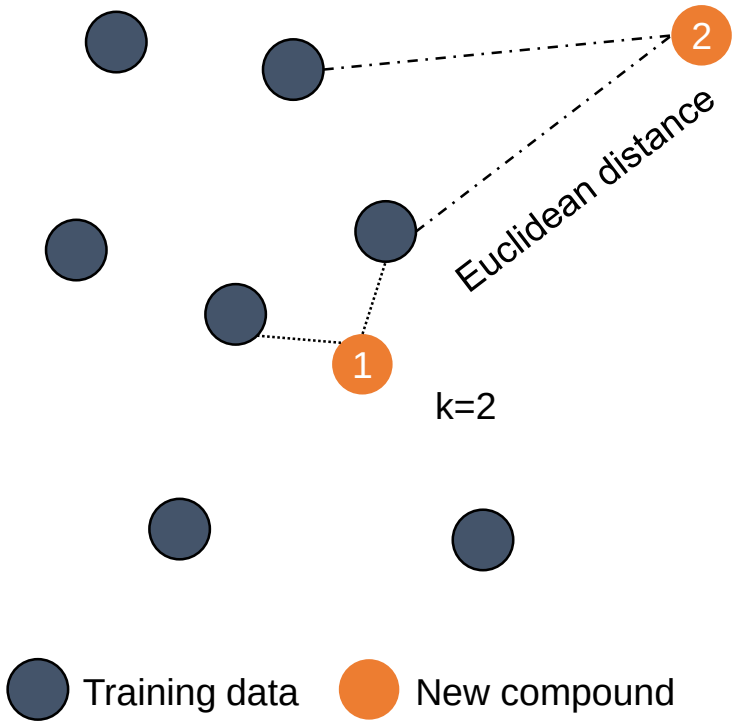
$$RI_{(x=1)} = \sum_{y=1}^{m=2} w_{xy} w_{yz}$$

$$= (-0.8604 * -5.2075) + (1.0651 * 2.0158) = 6.6276$$

Ibrahim, O. M. "A comparison of methods for assessing the relative importance of input variables in artificial neural networks." Journal of Applied Sciences Research 9.11 (2013): 5692-5700.

Model Interpretation : Applicability Domain

- **KNN-based applicability domain**

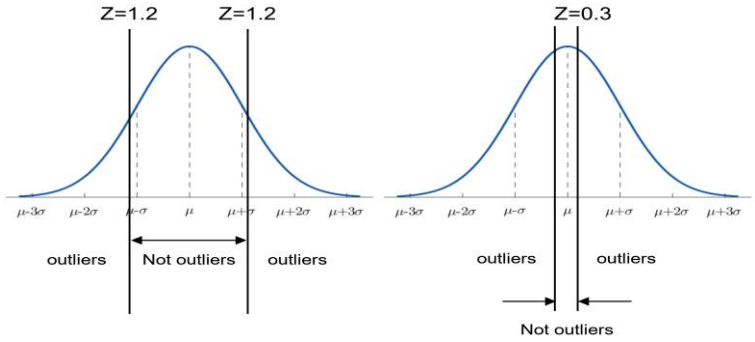


The new compound will be predicted by the model, only if:

$$\mathbf{D}_i \leq < \mathbf{D}_k > + \mathbf{Z} \times \mathbf{s}_k$$

With Z, an empirical parameter(0.5 by default)

$<\mathbf{D}_k>$: average Euclidian distance between each compound of the training set and its k nearest neighbors in the descriptors space.
$\mathbf{s}_k$: standard deviation of the distances between each compound of the training set and its k nearest neighbors in the descriptors space.
$\mathbf{D}_i$: the average of the distances between i and its k nearest neighbors in the training set.



Tropsha, Alexander, Paola Gramatica, and Vijay K. Gombar. "The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models." Molecular Informatics 22.1 (2003): 69-77.