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


How can we bridge the gap between Excellence in nano R&D and Profitable nano Industry?
Safe & Sustainable Nano-technology Portal:
A Toolbox for Nanomaterial Characterization & Safety Assessments
www.s2nano.org

http://portal.s2nano.org

“사용자 친화적 나노안전성 예측시스템”
Measurements to Models
For Nanomaterials

- Surface area
- Shape
- Size
- Surface chemistry
- COOH
- Surface charge
- OH
- NH3
- Agglomeration
- Dissolutions
- Dispersion stability
- Compositions & Impurities
- Concentrations in mass, number, & surface area

- Decision Making
- User-Friendly Interface
- Nanosafety Prediction Models
- Prediction Models
- Quality Screened Datasets
- s2NANO Database
- Raw Data: Experiments, literature, Report & Simulations
User Friendly Nanosafety Prediction System

1. Reference NM Library
2. Standard NM Library
3. Manufactured NM Library

PChem Data (Measured) → NanoToxicity Prediction Models → Safety Index

Tox Data (Measured) → Safety Index

Tox Data (Literature) → Safety Index

Input
- Manufactured Nanomaterials
- SiO₂
- 3mm ~ 300nm
- -30mV ~ +30mV
- Various properties

Output
- Safe
- Unsafe
- Safety Index

Reference NM Library
- Standard NM Library
- Manufactured NM Library

PChem Data (Literature)

Tox Data (Measured)

Tox Data (Literature)

Manufactured NM Library

Safety?

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 $S^2$NANO

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
  - www.s2nano.org
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

www.s2nano.org
Data Collection - Literature Mining

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
- 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),,
Improve Data Quality via Scoring Methods

Collected Database

Chemical Research in Toxicology
Quasi-SMILES-Based Nano-Quantitative Structure-Activity Relationship Model to Predict the Cytotoxicity of Multinwalled Carbon Nanotubes to Human Lung Cells

① Info
② Material
③ PChem
④ Toxicity
Model Development Workflow in S²NANO

Data Collection
- Literature Search
- Information Extraction
- Information Input in Nanosafety Database
- Data Selection from Nanosafety Database

Data Preprocessing
- Attribute Selection
- Data Normalization (Mean Centering and Scaling)
- Data Gap Filling
- Score-based Screening

Data Completeness (Missing Data Problem)
Data Quality (Heterogeneous Source of Data)
Data Imbalance (NonToxic >> Toxic)

Model Development
- Training/Test set Splitting (60% training set, 40% test set)
- Model Training (Random Forest Algorithm)
- Model Validation (Internal and External)

Knowledge Extraction
- Model Interpretation
- Applicability Domain
Data Preprocessing - Attribute Selection

- **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 $\Delta H_{sf}$ (eV)
  - Conduction band energy $E_c$ (eV)
  - Valence band energy $E_v$ (eV)
  - Electronegativity $\chi_{MeO}$ (eV)

Nanosafety database

www.s2nano.org
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 Endpoint
  - Toxic when Cell Viability < 50%
  - NonToxic when Cell Viability ≥ 50%

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

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

**Dataset I**
- Original dataset
- Mean substitution

- 6842 rows
  score = 2.8 ± 1.3

**Dataset II**
- Data gap filling

- 3246 rows
  score = 4.7 ± 0.2

**Dataset III**
- Data gap filling
  - 50% data with top PChem score

- 1738 rows
  score = 4.8 ± 0.1

**Dataset III**
- Data gap filling
  - 20% data with top PChem score

- 666 rows
  score = 4.9 ± 0.2
Data Imbalance Issue

- SMOTE (Synthetic minority over-sampling technique)


**Toxic 16% : Nontoxic 84 %**

**SMOTE (Synthetic Minority Over-sampling TEtchnique )**

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

PredictNANO > Datasets

Download Dataset

Details on Dataset
Model Development Workflow in $S^2\text{NANO}$

Data Collection
- Literature Search
- Information Extraction
- Information Input in Nanosafety Database
- Data Selection from Nanosafety Database

Data Preprocessing
- Attribute Selection
- Data Normalization
  - Mean Centering and Scaling
- Data Gap Filling
- Score-based Screening

Model Development
- Training/Test set Splitting
  - 60% training set, 40% test set
- Model Training
  - Random Forest Algorithm
- Model Validation
  - Internal and External

Knowledge Extraction
- Model Interpretation
- Applicability Domain

Logistic Regression Algorithm
Random Forest Algorithm
Support Vector Machine Algorithm
Backpropagation Algorithm
Model Development – Algorithm Selection

A. Random Forest Algorithm
B. Support Vector Machine Algorithm
C. Logistic Regression Algorithm
D. Backpropagation Algorithm
E. Radial Basis Function Network (RBFN) Algorithm
Model Development – Validation (Internal & External)

Model validation

<table>
<thead>
<tr>
<th>Validation</th>
<th>Dataset</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal validation</td>
<td>Training set (normalized by each method)</td>
<td>- 1. Select the normalization method appropriate for each modeling algorithm</td>
</tr>
<tr>
<td></td>
<td>Balanced training set (by SMOTE)</td>
<td>- 2. Look at SMOTE effect</td>
</tr>
<tr>
<td>External validation</td>
<td>Test set</td>
<td>- 1. Look at SMOTE effect</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- 2. Select the best predictive model</td>
</tr>
<tr>
<td>Reliability validation</td>
<td>Cytotoxicity data measured by experiment</td>
<td>- Validate the reliability of the best predictive model</td>
</tr>
</tbody>
</table>

Performance measures

- **True condition**
  - positive toxic: True Positive (TP)
  - negative nontoxic: False Negative (FN)

- **Predicted condition**
  - positive toxic
  - negative nontoxic

- **Sensitivity** = \( \frac{TP}{TP + FP} \)
- **Specificity** = \( \frac{TN}{FN + TN} \)

- **Accuracy** = \( \frac{TP + TN}{TP + FN + FP + TN} \)

- **Balanced accuracy** = \( \frac{1}{2} \left( \frac{TP}{TP + FP} + \frac{TN}{FN + TN} \right) \)

Choi et al. (2018)
### Normalization Method

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Normalization method</th>
<th>True positive</th>
<th>False positive</th>
<th>False negative</th>
<th>True negative</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Balanced accuracy</th>
<th>Standard deviation</th>
<th>Balanced accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>min-max</td>
<td>39</td>
<td>12</td>
<td>16</td>
<td>278</td>
<td>70.91%</td>
<td>95.86%</td>
<td>83.39%</td>
<td>4.84%</td>
<td>2.45%</td>
</tr>
<tr>
<td></td>
<td>z-score</td>
<td>39</td>
<td>12</td>
<td>16</td>
<td>278</td>
<td>70.91%</td>
<td>95.86%</td>
<td>83.39%</td>
<td>4.84%</td>
<td>2.47%</td>
</tr>
<tr>
<td></td>
<td>log</td>
<td>46</td>
<td>11</td>
<td>9</td>
<td>279</td>
<td>83.64%</td>
<td>96.21%</td>
<td>89.92%</td>
<td>4.45%</td>
<td>2.33%</td>
</tr>
<tr>
<td></td>
<td>combination</td>
<td>45</td>
<td>8</td>
<td>10</td>
<td>282</td>
<td>81.82%</td>
<td>97.24%</td>
<td>89.53%</td>
<td>3.99%</td>
<td>3.99%</td>
</tr>
<tr>
<td>SVM</td>
<td>min-max</td>
<td>28</td>
<td>6</td>
<td>27</td>
<td>284</td>
<td>50.91%</td>
<td>97.93%</td>
<td>74.42%</td>
<td>10.07%</td>
<td>5.04%</td>
</tr>
<tr>
<td></td>
<td>z-score</td>
<td>29</td>
<td>7</td>
<td>26</td>
<td>283</td>
<td>52.73%</td>
<td>97.59%</td>
<td>75.16%</td>
<td>9.73%</td>
<td>4.87%</td>
</tr>
<tr>
<td></td>
<td>log</td>
<td>40</td>
<td>5</td>
<td>15</td>
<td>285</td>
<td>72.73%</td>
<td>98.28%</td>
<td>85.50%</td>
<td>5.48%</td>
<td>2.65%</td>
</tr>
<tr>
<td></td>
<td>combination</td>
<td>41</td>
<td>5</td>
<td>14</td>
<td>285</td>
<td>74.55%</td>
<td>98.28%</td>
<td>86.41%</td>
<td>5.48%</td>
<td>2.73%</td>
</tr>
<tr>
<td>RF</td>
<td>min-max</td>
<td>45</td>
<td>5</td>
<td>10</td>
<td>285</td>
<td>81.82%</td>
<td>98.28%</td>
<td>90.05%</td>
<td>5.71%</td>
<td>2.92%</td>
</tr>
<tr>
<td></td>
<td>z-score</td>
<td>44</td>
<td>5</td>
<td>11</td>
<td>285</td>
<td>80.00%</td>
<td>98.28%</td>
<td>89.14%</td>
<td>4.32%</td>
<td>2.25%</td>
</tr>
<tr>
<td></td>
<td>log</td>
<td>45</td>
<td>5</td>
<td>10</td>
<td>285</td>
<td>81.82%</td>
<td>98.28%</td>
<td>90.05%</td>
<td>5.02%</td>
<td>2.49%</td>
</tr>
<tr>
<td></td>
<td>combination</td>
<td>45</td>
<td>5</td>
<td>10</td>
<td>285</td>
<td>81.82%</td>
<td>98.28%</td>
<td>90.05%</td>
<td>4.93%</td>
<td>2.44%</td>
</tr>
<tr>
<td>ANN</td>
<td>min-max</td>
<td>38</td>
<td>15</td>
<td>17</td>
<td>275</td>
<td>69.09%</td>
<td>94.83%</td>
<td>81.96%</td>
<td>15.31%</td>
<td>7.79%</td>
</tr>
<tr>
<td></td>
<td>z-score</td>
<td>40</td>
<td>6</td>
<td>15</td>
<td>284</td>
<td>72.73%</td>
<td>97.93%</td>
<td>85.33%</td>
<td>5.89%</td>
<td>2.98%</td>
</tr>
<tr>
<td></td>
<td>log</td>
<td>43</td>
<td>8</td>
<td>12</td>
<td>282</td>
<td>78.18%</td>
<td>97.24%</td>
<td>87.71%</td>
<td>6.00%</td>
<td>3.01%</td>
</tr>
<tr>
<td></td>
<td>combination</td>
<td>48</td>
<td>8</td>
<td>7</td>
<td>282</td>
<td><strong>87.27%</strong></td>
<td><strong>97.24%</strong></td>
<td><strong>92.26%</strong></td>
<td><strong>4.05%</strong></td>
<td><strong>2.11%</strong></td>
</tr>
<tr>
<td>Normalization method</td>
<td>Algorithm</td>
<td>Data</td>
<td>True positive</td>
<td>False positive</td>
<td>False negative</td>
<td>True negative</td>
<td>Sensitivity</td>
<td>Specificity</td>
<td>Balanced accuracy</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------</td>
<td>-------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>--------------</td>
<td>-------------</td>
<td>-------------</td>
<td>------------------</td>
<td>------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ID</td>
<td>46</td>
<td>11</td>
<td>9</td>
<td>279</td>
<td>83.64%</td>
<td>96.21%</td>
<td>89.92%</td>
<td>4.45%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BD</td>
<td>243</td>
<td>27</td>
<td>18</td>
<td>234</td>
<td>93.10%</td>
<td>89.66%</td>
<td>91.38%</td>
<td>5.84%</td>
</tr>
<tr>
<td>Log</td>
<td>LR</td>
<td>ID</td>
<td>41</td>
<td>5</td>
<td>14</td>
<td>285</td>
<td>74.55%</td>
<td>98.28%</td>
<td>86.41%</td>
<td>5.48%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BD</td>
<td>255</td>
<td>7</td>
<td>6</td>
<td>254</td>
<td>97.70%</td>
<td>97.32%</td>
<td>97.51%</td>
<td>1.15%</td>
</tr>
<tr>
<td>Combination</td>
<td>SVM</td>
<td>ID</td>
<td>45</td>
<td>5</td>
<td>10</td>
<td>285</td>
<td>81.82%</td>
<td>98.28%</td>
<td>90.05%</td>
<td>4.93%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BD</td>
<td>253</td>
<td>4</td>
<td>8</td>
<td>257</td>
<td>96.93%</td>
<td>98.47%</td>
<td>97.70%</td>
<td>0.88%</td>
</tr>
<tr>
<td>Combination</td>
<td>RF</td>
<td>ID</td>
<td>48</td>
<td>8</td>
<td>7</td>
<td>282</td>
<td>87.27%</td>
<td>97.24%</td>
<td>92.26%</td>
<td>4.05%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BD</td>
<td>258</td>
<td>5</td>
<td>3</td>
<td>256</td>
<td>98.85%</td>
<td>98.08%</td>
<td>98.47%</td>
<td>0.79%</td>
</tr>
</tbody>
</table>

Toxic 16% : Nontoxic 84 %
SMOTE (Synthetic Minority Over-sampling TTechnique )
ID (Imbalanced Data) vs. BD (Balanced Data)
<table>
<thead>
<tr>
<th>Normalization method</th>
<th>Algorithm</th>
<th>Data</th>
<th>True positive</th>
<th>False positive</th>
<th>False negative</th>
<th>True negative</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Balanced accuracy</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>退化</td>
<td>Log</td>
<td>LR</td>
<td>ID</td>
<td>46</td>
<td>11</td>
<td>9</td>
<td>279</td>
<td>83.64%</td>
<td>96.21%</td>
<td>89.92%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BD</td>
<td>243</td>
<td>27</td>
<td>18</td>
<td>187</td>
<td>234</td>
<td>93.10%</td>
<td>89.66%</td>
<td>91.38%</td>
</tr>
<tr>
<td>内部验证</td>
<td>Combination</td>
<td>SVM</td>
<td>ID</td>
<td>41</td>
<td>5</td>
<td>14</td>
<td>285</td>
<td>74.55%</td>
<td>98.28%</td>
<td>86.41%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BD</td>
<td>255</td>
<td>7</td>
<td>6</td>
<td>9</td>
<td>254</td>
<td>97.70%</td>
<td>97.32%</td>
<td>97.51%</td>
</tr>
<tr>
<td></td>
<td>Combination</td>
<td>RF</td>
<td>ID</td>
<td>45</td>
<td>5</td>
<td>10</td>
<td>285</td>
<td>81.82%</td>
<td>98.28%</td>
<td>90.05%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BD</td>
<td>253</td>
<td>4</td>
<td>8</td>
<td>13</td>
<td>257</td>
<td>96.93%</td>
<td>98.47%</td>
<td>97.70%</td>
</tr>
<tr>
<td></td>
<td>Combination</td>
<td>ANN</td>
<td>ID</td>
<td>48</td>
<td>8</td>
<td>7</td>
<td>282</td>
<td>87.27%</td>
<td>97.24%</td>
<td>92.26%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BD</td>
<td>258</td>
<td>5</td>
<td>3</td>
<td>13</td>
<td>256</td>
<td>98.85%</td>
<td>98.08%</td>
<td>98.47%</td>
</tr>
<tr>
<td>外部验证</td>
<td>Log</td>
<td>LR</td>
<td>ID</td>
<td>25</td>
<td>6</td>
<td>4</td>
<td>194</td>
<td>86.21%</td>
<td>97.00%</td>
<td>91.60%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BD</td>
<td>26</td>
<td>21</td>
<td>3</td>
<td>179</td>
<td>191</td>
<td>89.66%</td>
<td>89.50%</td>
<td>89.58%</td>
</tr>
<tr>
<td></td>
<td>Combination</td>
<td>SVM</td>
<td>ID</td>
<td>22</td>
<td>5</td>
<td>7</td>
<td>195</td>
<td>75.86%</td>
<td>97.50%</td>
<td>86.68%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BD</td>
<td>25</td>
<td>10</td>
<td>4</td>
<td>190</td>
<td>191</td>
<td>86.21%</td>
<td>95.00%</td>
<td>90.60%</td>
</tr>
<tr>
<td></td>
<td>Combination</td>
<td>RF</td>
<td>ID</td>
<td>24</td>
<td>3</td>
<td>5</td>
<td>197</td>
<td>82.76%</td>
<td>98.50%</td>
<td>90.63%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BD</td>
<td>25</td>
<td>9</td>
<td>4</td>
<td>191</td>
<td>191</td>
<td>86.21%</td>
<td>95.50%</td>
<td>90.85%</td>
</tr>
<tr>
<td></td>
<td>Combination</td>
<td>ANN</td>
<td>ID</td>
<td>23</td>
<td>4</td>
<td>6</td>
<td>196</td>
<td>79.31%</td>
<td>98.00%</td>
<td>88.66%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BD</td>
<td>27</td>
<td>13</td>
<td>2</td>
<td>187</td>
<td>200</td>
<td>93.10%</td>
<td>93.50%</td>
<td>93.30%</td>
</tr>
</tbody>
</table>

**Internal vs. External Validations**

<table>
<thead>
<tr>
<th>Normalization method</th>
<th>Algorithm</th>
<th>Data</th>
<th>True positive</th>
<th>False positive</th>
<th>False negative</th>
<th>True negative</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Balanced accuracy</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>退化</td>
<td>Log</td>
<td>LR</td>
<td>ID</td>
<td>25</td>
<td>6</td>
<td>4</td>
<td>194</td>
<td>86.21%</td>
<td>97.00%</td>
<td>91.60%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BD</td>
<td>26</td>
<td>21</td>
<td>3</td>
<td>179</td>
<td>191</td>
<td>89.66%</td>
<td>89.50%</td>
<td>89.58%</td>
</tr>
<tr>
<td>内部验证</td>
<td>Combination</td>
<td>SVM</td>
<td>ID</td>
<td>22</td>
<td>5</td>
<td>7</td>
<td>195</td>
<td>75.86%</td>
<td>97.50%</td>
<td>86.68%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BD</td>
<td>25</td>
<td>10</td>
<td>4</td>
<td>190</td>
<td>191</td>
<td>86.21%</td>
<td>95.00%</td>
<td>90.60%</td>
</tr>
<tr>
<td></td>
<td>Combination</td>
<td>RF</td>
<td>ID</td>
<td>24</td>
<td>3</td>
<td>5</td>
<td>197</td>
<td>82.76%</td>
<td>98.50%</td>
<td>90.63%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BD</td>
<td>25</td>
<td>9</td>
<td>4</td>
<td>191</td>
<td>191</td>
<td>86.21%</td>
<td>95.50%</td>
<td>90.85%</td>
</tr>
<tr>
<td></td>
<td>Combination</td>
<td>ANN</td>
<td>ID</td>
<td>23</td>
<td>4</td>
<td>6</td>
<td>196</td>
<td>79.31%</td>
<td>98.00%</td>
<td>88.66%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BD</td>
<td>27</td>
<td>13</td>
<td>2</td>
<td>187</td>
<td>200</td>
<td>93.10%</td>
<td>93.50%</td>
<td>93.30%</td>
</tr>
</tbody>
</table>
Performance Comparisons of nanoSAR classification Models

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Dataset</th>
<th>Algorithm</th>
<th>Internal validation</th>
<th>External validation</th>
<th>Publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>PM-100</td>
<td>2017_Metal_HYU</td>
<td>Random forest</td>
<td>86%</td>
<td>82%</td>
<td>Ha et al. (2018) Scientific Reports</td>
</tr>
<tr>
<td>PM-101</td>
<td>2016_Metal_KNU</td>
<td>Backpropagation</td>
<td>95%</td>
<td>67%</td>
<td>Choi et al. (2018) Scientific Reports</td>
</tr>
<tr>
<td>PM-102</td>
<td>2015_Metal_HYU</td>
<td>Support vector machine</td>
<td>90%</td>
<td>75%</td>
<td></td>
</tr>
<tr>
<td>PM-103</td>
<td>2015_Metal_MeOx_KNU</td>
<td>Backpropagation</td>
<td>91%</td>
<td>86%</td>
<td></td>
</tr>
<tr>
<td>PM-104</td>
<td>2015_Metal_HYU</td>
<td>Support vector machine</td>
<td>90%</td>
<td>82%</td>
<td></td>
</tr>
<tr>
<td>PM-106</td>
<td>2017_MeOx_I_KNU</td>
<td>Resilient backpropagation</td>
<td>93%</td>
<td>86%</td>
<td></td>
</tr>
<tr>
<td>PM-107</td>
<td>2017_MeOx_IL_HYU</td>
<td>Random forest</td>
<td>94%</td>
<td>54%</td>
<td>Ha et al. (2018) Scientific Reports</td>
</tr>
<tr>
<td>PM-108</td>
<td>2016_MeOx_KNU</td>
<td>Backpropagation</td>
<td>95%</td>
<td>85%</td>
<td>Choi et al. (2018) Scientific Reports</td>
</tr>
<tr>
<td>PM-114</td>
<td>2017_MeOx_I_KNU</td>
<td>Random forest</td>
<td>91%</td>
<td>87%</td>
<td></td>
</tr>
<tr>
<td>PM-115</td>
<td>2017_MeOx_I_KNU</td>
<td>Support vector machine</td>
<td>93%</td>
<td>89%</td>
<td>Trinh et al. (2018) Chemosphere</td>
</tr>
<tr>
<td>PM-116</td>
<td>2017_MWCNT_HYU</td>
<td>Quasi-QSAR</td>
<td>0.89</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>PM-117</td>
<td>2017_MeOx_IL_KNU</td>
<td>Quasi-QSAR</td>
<td>0.79</td>
<td>0.71</td>
<td>Choi et al. (2018) Chemosphere</td>
</tr>
<tr>
<td>PM-118</td>
<td>2015_Metal_KNU</td>
<td>Backpropagation</td>
<td>88%</td>
<td>67%</td>
<td></td>
</tr>
<tr>
<td>PM-120</td>
<td>2017_Metal_HYU</td>
<td>Support vector machine</td>
<td>86%</td>
<td>82%</td>
<td>Trinh et al. (2018) Environmental Science : NANO</td>
</tr>
</tbody>
</table>

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

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

PredictNANO > Models

Toxicity Prediction Datawarehouse

Details on Each Prediction Models
Go "Safety Screening"

Material Group: Oxides

Material: ZnO

Core Size

Hydrodynamic Size

Surface Charge

Mass Dose
Model Development Workflow in S²NANO

Data Collection
- Literature Search
- Information Extraction
- Information Input in Nanosafety Database
- Data Selection from Nanosafety Database

Data Preprocessing
- Attribute Selection
- Data Normalization
- Mean Centering and Scaling
- Data Gap Filling
- Score-based Screening

Model Development
- Training/Test set Splitting
  60% training set, 40% test set
- Model Training
  Random Forest Algorithm
- Model Validation
  Internal and External

Knowledge Extraction
- Model Interpretation
- Applicability Domain

Relative Importance of Attributes
Applicability Domains
Figure 4. Leave-one-out OOB errors against attributes.

Table 8. Relative importance.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Relative importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dose</td>
<td>11.10</td>
</tr>
<tr>
<td>ΔHsf</td>
<td>7.34</td>
</tr>
<tr>
<td>Exposure time</td>
<td>5.33</td>
</tr>
<tr>
<td>Hydrodynamic size</td>
<td>4.43</td>
</tr>
<tr>
<td>Ec</td>
<td>3.66</td>
</tr>
<tr>
<td>Surface area</td>
<td>3.58</td>
</tr>
<tr>
<td>Core size</td>
<td>3.57</td>
</tr>
<tr>
<td>Cell species</td>
<td>2.88</td>
</tr>
<tr>
<td>xMeO</td>
<td>2.44</td>
</tr>
<tr>
<td>Cell type</td>
<td>2.44</td>
</tr>
<tr>
<td>Surface charge</td>
<td>1.90</td>
</tr>
<tr>
<td>Assay method</td>
<td>1.82</td>
</tr>
<tr>
<td>Ev</td>
<td>1.45</td>
</tr>
<tr>
<td>Cell name</td>
<td>1.33</td>
</tr>
<tr>
<td>Cell origin</td>
<td>1.06</td>
</tr>
</tbody>
</table>
**Applicability Domains of Models**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>I</th>
<th>II</th>
<th>III-A</th>
<th>III-B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min.</td>
<td>Max.</td>
<td>Min.</td>
<td>Max.</td>
</tr>
<tr>
<td>Dose (µg/mL)</td>
<td>0</td>
<td>10000</td>
<td>0</td>
<td>167000</td>
</tr>
<tr>
<td>Time (h)</td>
<td>0</td>
<td>360</td>
<td>1</td>
<td>168</td>
</tr>
<tr>
<td>Core size (nm)</td>
<td>2.7</td>
<td>629</td>
<td>2.7</td>
<td>496</td>
</tr>
<tr>
<td>Hydro. size (nm)</td>
<td>8.6</td>
<td>6181</td>
<td>8.6</td>
<td>2300</td>
</tr>
<tr>
<td>Surface charge (mV)</td>
<td>-63.3</td>
<td>61.9</td>
<td>-63.3</td>
<td>61.9</td>
</tr>
<tr>
<td>Surface area (m²/g)</td>
<td>0.8</td>
<td>1150</td>
<td>5.5</td>
<td>576</td>
</tr>
<tr>
<td>ΔΗₐ (eV)</td>
<td>-64.7</td>
<td>-1.2</td>
<td>-64.7</td>
<td>-1.2</td>
</tr>
<tr>
<td>Eₐ (eV)</td>
<td>-6.6</td>
<td>-0.1</td>
<td>-6.6</td>
<td>-0.1</td>
</tr>
<tr>
<td>Eₑ (eV)</td>
<td>-11.4</td>
<td>-5.0</td>
<td>-11.3</td>
<td>-5.0</td>
</tr>
<tr>
<td>Χ (eV)</td>
<td>3.2</td>
<td>8.3</td>
<td>3.4</td>
<td>8.3</td>
</tr>
</tbody>
</table>

Table 4. Applicability domains regarding the numerical attributes.

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

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Dataset A</th>
<th>Dataset B</th>
<th>Dataset C</th>
</tr>
</thead>
<tbody>
<tr>
<td>NPs type</td>
<td>Ag, Au</td>
<td>Ag, Au</td>
<td>Ag, Au</td>
</tr>
<tr>
<td>Shape</td>
<td>Particle, hollow, nanorod</td>
<td>Particle, hollow, nanorod</td>
<td>Particle, hollow, nanorod</td>
</tr>
<tr>
<td>Core size (nm)</td>
<td>2 – 120</td>
<td>2 – 120</td>
<td>2.5 – 120</td>
</tr>
<tr>
<td>Hydrodynamic size (nm)</td>
<td>7.1 – 300</td>
<td>7.1 – 300</td>
<td>7.1 – 300</td>
</tr>
<tr>
<td>Surface charge (mV)</td>
<td>-78.8 – 58.2</td>
<td>-78.8 – 58.2</td>
<td>-78.8 – 58.2</td>
</tr>
<tr>
<td>Specific surface area (m²/g)</td>
<td>2.3 – 185.7</td>
<td>2.3 – 185.7</td>
<td>2.3</td>
</tr>
<tr>
<td>Dose (ppm)</td>
<td>0 – 400</td>
<td>0 – 400</td>
<td>0 – 400</td>
</tr>
<tr>
<td>Exposure time (h)</td>
<td>0 – 96</td>
<td>0 – 96</td>
<td>0 – 96</td>
</tr>
</tbody>
</table>
Datawarehouse with User-Friendly Interfaces

Description of Screening Models

Other safety indexes

No Decision

Model result

Performance

Sensitivity

Accuracy

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

Raw Data = 33,393 rows

Quality Screened Datasets = 16

Prediction Models = 13

User-Friendly Interface

S²NANO Database

Nanosafety Prediction Models

Raw Data: Experiments, literature, Report & Simulations

Decision Making
Excellence in nano R&D

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 design and regulation compliance.

Profitable nano Industry
Measurements & Models for Nanomaterials

NanoSolveIT
Nanoprediction Toolbox

Predictive Models
In the case of “classic” QSPR/QSAR analysis the paradigm is the following:

\[ \text{Endpoint} = \text{Mathematical function (Molecular structure)} \]

In the case of nanomaterials, the molecular structures of nanomaterials is the same as bulk chemicals.

**Problem** In the case of nanomaterials, the molecular structures of nanomaterials is the same as bulk chemicals.

**Solution** replace the traditional paradigm with by fresh paradigm

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


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: 0.46-0.50)

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

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)
Thanks for your attention
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

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

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

\[ = (-0.8604 \times -5.2075) + (1.0651 \times 2.0158) = 6.6276 \]

KNN-based applicability domain

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

\[ D_i \leq <D_k> + Z \times s_k \]

With Z, an empirical parameter (0.5 by default)

\(<D_k>\) : average Euclidian distance between each compound of the training set and its k nearest neighbors in the descriptors space.

\(s_k\) : standard deviation of the distances between each compound of the training set and its k nearest neighbors in the descriptors space.

\(D_i\) : the average of the distances between \(i\) and its k nearest neighbors in the training set.