Predicting Morphological Profiles Associated with DNA Damage or Oxidative Stress Using Artificial Intelligence Models

Chaeyoung Seo[1] and Seung Jin Lee[1,2,*]

[1]*Department of Bio-AI Convergence*

(AI-Biology & Pharmaceutical Science) Chungnam University

[2]*Department of Pharmacology, College of Pharmacy, Chungnam University*

**Corresponding author: s.j.lee@cnu.ac.kr*

The Cell Painting assay is a phenotypic profiling technique to capture cellular response to perturbagens. Artificial intelligence (AI) models can predict various aspects of cell health with cell painting features, but some of cell health outcomes have been predicted with limited accuracy. Our study aimed to develop a high-performance AI model to predict cell phenotypes associated with DNA damage or reactive oxygen species (ROS) production, utilizing the idr-0080 dataset. Additionally, this model was applied to cpg0012 dataset and outcomes were validated with literatures. For AI model training, features for DNA damage or ROS were visualized using Gaussian distribution and t-SNE, clustered with K-means and the elbow method, and labeled as either 'high' or 'low' class. A total of seven AI-models were trained, including Random Forest, Gradient Boosting, SVM, Logistic Regression, SGD Classifier, K-Neighbors, Gaussian NB, and Decision Tree, along with Deep Learning model. To overcome class imbalance and improve model performance on the minority class, each model was evaluated under four conditions: (1) without any modifications, (2) with oversampling using SMOTE, (3) with sample weights, and (4) with both oversampling and sample weights. The models' performance was assessed using F1-score, confusion matrix, and ROC curve analysis. Specifically, the DNA damage-predicting model, implemented with a deep learning architecture without additional balancing techniques, achieved an AUC score of 0.83. The top-performing model to predict ROS generation was a deep learning approach utilizing sample weights, yielding an AUC score of 0.75. When applied to the cpg0012 dataset, this model identified potential DNA damage- or ROS-inducing compounds, which were distinctively clustered by UMAP analysis. Of the 14 compounds with DNA damage features, 42.9%

were reported to induce DNA damage, and 50.0% were referred to induce cytotoxicity. Similarly, among 64 compounds with ROS-related features, 59.4% were linked to ROS production, and 25.0% to cytotoxicity in literatures. In conclusion, we developed high-performance AI models to predict DNA damage or oxidative stress using morphological profiles, highlighting their potential in discovering compounds that modulate cellular stress responses.