**A Machine Learning-Based Drug-Excipient Compatibility Prediction Model in Homogeneous Pharmaceuticals Based on Structural Interactions**

**Jun Hak Lee1**, Kwan Hyung Cho1,2.

Department of Pharmacy, Inje University1, Gimhae, 50834, Republic of Korea.

**Background and aims.** Appropriate excipient selection is crucial in the development of homogeneous dosage forms; however, conventional compatibility test remains labor and resource-intensive. Moreover, excipients can decrease drug quality by incompatibility of active pharmaceutical ingredients.

**Methods.**  A dataset of compatible drug–excipient pairs was setup from 356 products approved by the Korean Ministry of Food and Drug Safety (MFDS) and 79 products listed by the U.S. FDA, with all contained combinations labeled as compatible. Incompatible pairs were implemented from peer-reviewed publications since 2000 and the Handbook of Pharmaceutical Excipients. SMILES of drugs and excipients were standardized (removal of salts, solvates, hydrates) and converted into 1,024-bit molecular fingerprints via RDKit. Pair-level feature vectors comprising two fingerprints plus compatbility label—yielded 2,049-dimensional inputs. Machine learning models, including logistic regression and a multilayer perceptron (MLP), were implemented in scikit-learn. Performance of model was evaluated by accuracy, Matthews correlation coefficient (MCC), and area under the ROC curve (AUC). Prediction test was performed with a separately prepared data set. The prediction accuracy of the model was evaluated by considering the probability as compatibile if it was 0.5 or higher and incompatible if it was less than 0.5.

**Results.** Performance of logistic model was highest among all models (accuracy 0.913, AUC 0.904, MCC 0.666). On prediction test set, logistic regression and MLP reached prediction accuracy of 69.0% and 60.0%, respectively, but both models exhibited poor sensitivity to incompatibility cases. After retraining on the expanded dataset, prediction accuracy improved to 85.0% for logistic regression and 90.0% for MLP, demonstrating that continuous data integration enhances predictive performance.

**Conclusion/Discussion.** Machine learning models can accurately predict drug–excipient compatibility, thereby reducing labor and resource-intensive experiments. Continuous enhancement of new compatibility data further improves model robustness, facilitating more efficient development of homogeneous formulation.

**Acknowledgement.** This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (Grant number: NRF-2022R1A2C1003070).

**References:**

(1) Patel, S. et al. (2023) Int. J. Pharm. 637

(2) Hang, Nguyen Thu. et al. (2024) Int. J. Pharm. 653

(3) Chadha, R. et al. (2014) J. Pharm. Biomed. Anal. 87:82-97

(4) Wang, N. et al. (2021) Int. J. Pharm. 607