**Development of a Machine Learning-Based Predictive Model for Excipient Selection in Lipid-Based Formulations**

**Sung Mo Park**, Kwan Hyung Cho.

Department of Pharmacy, Inje University, Gimhae, Gyeongsangnam-do, Republic of Korea.

**Background and aims.** Lipid-based formulations (LBFs) have been used to improve the solubility of poorly water-soluble drugs. However, selecting appropriate oils, surfactants, and co-surfactants remains challenging due to unpredictable drug solubility in various excipients. This study aimed to develop a predictive machine learning (ML) model based on drug-excipient solubility data in LBFs.

**Methods.** A dataset of 1,127 drug-excipient pairs was constructed from 369 sources, including articles and commercial LBF products. Excipient types were labelled as carrier or anti-carrier, with anti-carriers defined as excipients exhibiting drug solubility ≤25% compared to carriers. Each substance was presented as a SMILES string. Data were converted into fingerprints and divided into training and validation sets with an 80:20 ratio. ML models such as logit, KNN, naive\_bayes, SVM, decision\_tree, random\_forest, MLP, and XGBoost were trained using 80% of the dataset and validated on the remaining 20%. The models were subsequently tested on an independent set excluded from training, to predict solubility classification (soluble: >0.5; insoluble ≤0.5). Model performance was evaluated using accuracy, AUC, and MCC metrics.

**Results.** Performance on the validation set showed that the logit model demonstrated the highest performance with accuracy, AUC, and MCC values of 0.784 ± 0.037, 0.756 ± 0.027, and 0.310 ± 0.114, respectively. On the independent test set, logit and XGBoost achieved 60% and 70% accuracy, although prediction performance for samples classified as insoluble remained limited. After retraining with the independent test set, accuracy improved to 85% and 90% for the logit and XGBoost models, respectively.

**Conclusion/Discussion.** The ML models predict drug solubility classification in LBFs, providing insight for excipient selection while reducing empirical screening. However, their limited prediction accuracy for insoluble cases indicates the need for reinforced training data. Continued learning with more data will enhance performance and expand model applicability.

**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) Bannigan, P. et al (2021) Advanced Drug Delivery Reviews 175:113806

(2) Murray, J. et al (2023) European Journal of Pharmaceutical Sciences 191:106562

(3) Hang, N. et al (2024) International Journal of Pharmaceutics 653:123884