Contribution ID: **8** Type: **Oral Presentation**

Comparative Analysis of Pre-trained Language Models for Chemical Toxicity Prediction

Thursday, 20 March 2025 16:20 (20 minutes)

The convergence of Natural Language Processing (NLP) and cheminformatics represents a groundbreaking approach to drug development, particularly in the critical domain of toxicity prediction. Early identification of toxic compounds is paramount in pharmaceutical research, as late-stage toxicity discoveries lead to substantial financial setbacks and delays market approval. While traditional methods often require extensive laboratory testing, the representation of chemical compounds as SMILES (Simplified Molecular Input Line Entry System) strings offers an innovative pathway for applying NLP techniques to chemical structure analysis, potentially revolutionizing the efficiency and accuracy of toxicity assessments.

Leveraging the computational resources provided by Academia Sinica Grid Computing (ASGC), this study comprehensively evaluates the efficacy of pre-trained language models for molecular toxicity prediction. We assess both transformer-based models (RoBERTa and ChemBERTa) and large language models (GPT-3.5 Turbo, GPT-4 Turbo, and GPT-4o) across three established benchmark datasets: ClinTox, Tox21, and ToxCast. Through systematic optimization, we demonstrate that while RoBERTa achieved exceptional performance on ClinTox (0.9898 F1-score), it failed to generalize effectively to other datasets. Conversely, ChemBERTa exhibited robust cross-dataset performance, maintaining remarkable F1-scores results across all three benchmarks (ClinTox: 0.9034, Tox21: 0.8232, ToxCast: 0.8162). Notably, LLMs demonstrated remarkable adaptability in few-shot learning scenarios, with GPT-3.5 Turbo achieving near-perfect performance on ToxCast (0.9923 F1-score).

Our findings reveal the transformative potential of integrating NLP techniques into toxicity prediction workflows. The superior performance of LLMs, particularly in scenarios with limited training data, suggests a paradigm shift in computational toxicology. This research establishes a foundation for more efficient and accurate early-stage drug discovery processes, potentially accelerating the development of safer therapeutic compounds while reducing development costs.

Primary authors: Prof. TUNG, Chun-Wei (Institute of Biotechnology and Pharmaceutical Research, National Health Research, Miaoli, Taiwan); Mr CHUNG, Yueh-Hsi (Department of Industrial Education, National Taiwan Normal University, Taipei, Taiwan); Prof. CHANG, Yung-Chun (Graduate Institute of Data Science, Taipei Medical University, Taipei, Taiwan)

Presenter: Mr CHUNG, Yueh-Hsi (Department of Industrial Education, National Taiwan Normal University, Taipei, Taiwan)

Session Classification: Artificial Intelligence (AI) - II

Track Classification: Track 10: Artificial Intelligence (AI)