Integrating Bioactivity and Sequence Features for Neural-Relevant Compound-Target Prediction

Eva Viesi*
Department of Computer Science, University of Verona, Verona, Italy
NBFC, National Biodiversity Future Center, Palermo, Italy

Rosalba Giugno
Department of Computer Science, University of Verona, Verona, Italy
NBFC, National Biodiversity Future Center, Palermo, Italy

Abstract:

Objective: Understanding how small molecules—including environmental pollutants—interact with biological systems is critical for advancing models of neural toxicity, computational neuroscience, and synthetic biological systems. This study presents a chemogenomic approach that leverages machine learning to predict compound-target interactions (CTIs), with a focus on air pollutants and their potential neurotoxicological effects. By integrating compound bioactivity descriptors and target sequence features, our methodology bridges computational modeling with complex biological signaling, including pathways relevant to neural health and disease.

Methods: We curated a dataset of 1,830 small-molecule air pollutants from the U.S. Environmental Protection Agency and linked them to their known biological targets using PubChem BioAssay [1]. Bioactivity signatures were generated using four Signaturizer models [2,3], and target sequence descriptors were computed using the iFeature package [4]. These representations were integrated into a unified feature space to perform supervised classification of CTIs. To address data imbalance, a One-Class Support Vector Machine (OCSVM) was used for negative sample selection [5]. Four machine learning classifiers were trained and assessed via nested cross-validation on five independently generated datasets.

Results: OCSVM-based sampling significantly outperformed random strategies, enhancing classifier performance across all metrics. The best-performing model demonstrated strong generalization to unseen CTIs, particularly those related to neural targets of air pollutants. Furthermore, molecular features derived from our curated pollutant dataset showed consistent recall performance, highlighting their robustness for downstream neurotoxicity prediction.

Conclusion: This work demonstrates the potential of integrating chemical, bioactivity-based, and sequence-based signatures for the computational prediction of biologically meaningful CTIs. The methodology supports scalable toxicity screening and can be adapted to study the molecular mechanisms underlying neurodegenerative disorders, offering insights applicable to both neural engineering and biologically inspired robotic systems.