TITLE: Application of Molecular Descriptors in the In Vivo Genotoxicity Prediction Using Machine Learning

AUTHORS: Iuri Barbosa Pereira[1]; Mauricio Homem de Mello[1]; Rogerio Salvini[2]
INSTITUTIONS: [1] University of Brasília – Brasília – DF, Brazil; [2] Federal University of Goiás – Goiânia – GO, Brazil

INTRODUCTION: Genotoxicity represents a significant public health concern, being associated with mutations and genomic instability. *In vivo* assays are considered more representative of biological complexity and are essential for predictive risk assessment. The use of molecular descriptors in computational models offers an efficient alternative to traditional experimental testing, contributing to savings in both time and resources. **OBJECTIVE:** Evaluate the effectiveness of classical molecular descriptors in predicting *in vivo* genotoxicity using machine learning algorithms and robust performance metrics. **MATERIALS AND METHODS:** A total of 2,223 substances with *in vivo* results were selected from public databases (GENE-TOX, CCRIS, and ECHA). Compounds were represented by SMILES strings, with structures verified and harmonized. A total of 366 molecular descriptors were calculated using the RDKit library, encompassing topological indices, electronic properties, structural complexity metrics, and features related to three-dimensional conformation. Additionally, structural alerts recognized by regulatory agencies were identified through substructure matching. Models were trained and evaluated using stratified 10-fold cross-validation, employing Random Forest, ExtraTrees, and LGBMClassifier algorithms. The primary metric was the F1-score, complemented by accuracy and ROC-AUC. **RESULTS AND CONCLUSION:** Models trained exclusively with *in vivo* data and using only molecular descriptors demonstrated consistent performance. The ExtraTreesClassifier model achieved an F1-score of 80.4%, accuracy of 80.6%, and ROC-AUC of 86.0%. The molecular descriptors alone proved highly effective in predicting *in vivo* genotoxicity. These findings highlight the relevance of *in vivo* data and classical descriptors as reliable tools in predictive toxicity models, supporting advances in the screening of safe compounds for pharmaceutical and environmental applications.

Keywords: Genotoxicity; Computational toxicology; Machine learning.