

MACHINE LEARNING WORKFLOW FOR ORGAN TOXICITY PREDICTION

1. Select target organ toxicity (β)
 - Identify target organ toxicities (β) from ToxRefDB
- 1.1. Identify chemicals associated with β
 - Find the positive chemicals (I_{β}^{+}), negative chemicals (I_{β}^{-}), and total chemicals ($I_{\beta} = I_{\beta}^{+} \cup I_{\beta}^{-}$)
 - $n_{\beta}^{+} = \text{number}(I_{\beta}^{+})$, $n_{\beta}^{-} = \text{number}(I_{\beta}^{-})$ and $n_{\beta} = n_{\beta}^{+} + n_{\beta}^{-}$
 - Identify β where $n_{\beta}^{+} \geq 50$ and $n_{\beta}^{-} \geq 50$
 - Construct ($n_{\beta} \times 1$) binary vector X^{β} to represent positive (1) and negative (0) chemicals
- 1.2. Obtain data for chemicals associated with β
 - For descriptor type, $\alpha \in \{\text{chm, bio, ct, bc, bct}\}$ construct data matrices X^{α} for I_{β} chemicals
 - Construct $X^{\alpha, \beta}$ by merging X^{α} and X^{β} using unique chemical identifiers in I_{β}
2. Predict and evaluate toxicity using supervised machine learning
 - For each $\alpha \in \{\text{chm, bio, ct, bc, bct}\}$:
 - For $n_i = 50$ to $\min(n_{\beta}^{+}, n_{\beta}^{-})$ with stepsize of 10:
 - Construct balanced subsets of $X^{\alpha, \beta}$
 - Repeat 10 times:
 - I_i^{+}, I_i^{-} = random subset of n_i chemicals from $I_{\beta}^{+}, I_{\beta}^{-}$ and $I_i = I_i^{+} \cup I_i^{-}$
 - $X_j^{\alpha, \beta} = X^{\alpha, \beta}[I_i]$ i.e. $X_j^{\alpha, \beta} \subset X^{\alpha, \beta}$
 - Vary number of descriptors (n_{ds}) from 5 to 25:
 - Repeat 10 times:
 - Conduct 5-fold cross-validation testing:
 - Split $X_j^{\alpha, \beta}$ into $\{X_{j,1}^{\alpha, \beta}, X_{j,2}^{\alpha, \beta}, X_{j,3}^{\alpha, \beta}, X_{j,4}^{\alpha, \beta}, X_{j,5}^{\alpha, \beta}\}$ balanced subsets
 - For $k \in \{1, 2, 3, 4, 5\}$:
 - $X_{\text{train}} = X_{j,m}^{\alpha, \beta} \cup X_{j,m'}^{\alpha, \beta} \cup X_{j,m''}^{\alpha, \beta} \cup X_{j,m'''}^{\alpha, \beta}$ where $k \notin \{m, \dots, m'''\}$
 - $X_{\text{test}} = X_{j,k}^{\alpha, \beta}$
 - Build Classifier (C) using X_{train} with top n_{ds} using ANOVA F-value
 - Test C using X_{test}
 - Save the performance scores for $\{\beta, \alpha, n_i, n_{ds}, C\}$