QSRR Modelling Procedure

The QSRR prediction workflow implemented in this study comprises the following steps:

1. **Calculation of molecular descriptors:**
   For each molecule, descriptors are computed from SMILES strings using a custom function based on the RDKit library. The resulting descriptors are structured into a DataFrame.

2. **Data filtering and cleaning:**
   Only molecules with complete descriptor values and an associated target value (retention index, RI) are retained. Columns containing missing or infinite values are removed. In addition, constant descriptors (zero variance) are excluded to eliminate non-informative variables.

3. **Data standardization:**
   Descriptor values are standardized (z-score) using StandardScaler from scikit-learn, ensuring consistent scaling across variables.

4. **Feature selection:**
   Univariate feature selection is performed using the SelectKBest method with the f_regression scoring function. The top *k* descriptors most correlated with the target variable are retained, with a default limit set to *k* = 160.

5. **Model construction and training:**
   Four regression algorithms are supported:

   - Ordinary linear regression (LinearRegression)
   - Ridge regression (Ridge)
   - Partial least squares regression (PLSRegression)
   - Random forest regression

The model is trained on a randomly defined training set (80%), with the remaining 20% used for testing, repeated at each iteration.

6. **Cross-validation:**
   A 6-fold cross-validation (KFold, *n* = 6) is performed at each iteration to estimate the model's generalization performance, based on the coefficient of determination (R^2).

7. **Performance evaluation:**
   At each iteration, the mean relative error (%) is computed on the test set (20%), based on the absolute difference between predicted and observed RI values.

8. **Prediction of new molecules:**
   For each query molecule, the trained model predicts the RI value. This process is repeated over *n* = 500 independent iterations. The mean and standard deviation of predicted RI values are then reported to estimate prediction uncertainty.