Global Model:

Rat oral LD_{50} is predicted by a linear equation, summing the occurrences of used fragmental descriptors (f_i) multiplied by their statistical coefficients (a_i), the intercept (c), and the remaining unexplained variation (ε):

$$\log LD_{50} = \sum a_i f_i + c + \varepsilon$$

The bootstrapping method implies random compound sampling from the initial training set, i.e. generation of new 'training sub-sets' and derivation of an independent PLS model for each sub-set. This procedure is performed 100 times. Each of the sampled sub-sets is of the same size as the initial training set; however, the random manner of their population results in some compounds being selected more than once, and others being omitted. Therefore, the global QSAR model represents an ensemble of 100 PLS models, providing each compound with a vector of 100 LD_{50} predictions, each based on a slightly different sub-set of the initial training set. The global (or baseline) prediction is derived as a mean average of all 100 LD_{50} predictions.

These LD_{50} prediction vectors are then used to determine the similarity between any given pair of compounds.

Local corrections (Δ)

Each global prediction is subjected to the local similarity correction procedure, in order to capture the deviations from linear trend of the responses that may occur in specific chemical spaces.

The correction procedure is based on the analysis of the performance of the global PLS QSAR model in the local environment of the query compound. This means a comparison of the experimental data and PLS baseline predictions for the five most similar compounds from the training set. If baseline predictions for these compounds show any systematic deviations from their reported measured values, a local correction is applied to the LD_{50} baseline prediction of the query compound. The required correction (Δ) is calculated as a weighted average from the differences between global QSAR predictions and experimental data for the five most similar compounds in the training set:

$$\Delta = \sum_{i=1}^{n} a^{i-1} \cdot SI_i \cdot \Delta_i / \sum_{i=1}^{n} a^{i-1}.$$

where,

Δ correction that should be applied for the given prediction from the global model;

α a constant, influencing calculation of the weighted average, the simple average value will be calculated if this constant is set to 1;

SI_i similarity (individual Similarity Index) between given compound and the i-th most similar compound in the training set, calculated as the correlation coefficient between corresponding vectors, made of multiple estimated values from bootstrapping PLS models;

Δ_i the difference between the measured value and the value predicted by the global model for the i-th most similar compound: $\Delta_i = Yi - \hat{Y}$;

n a constant, that determines how many similar compounds should be taken into consideration while estimating correction.