Modeling framework

An essential part of the OCHEM platform is the modeling framework. Its main purpose is to provide facilities for the development of predictive computational models for physicochemical and biological properties of compounds. The framework is integrated with the database of experimental data and includes all the necessary steps required to build a computational model: data preparation, calculation and filtering of molecular descriptors, application of machine learning methods and analysis of a models' performance. This section gives an overview of these features and of the steps required to build a computational model in the OCHEM.

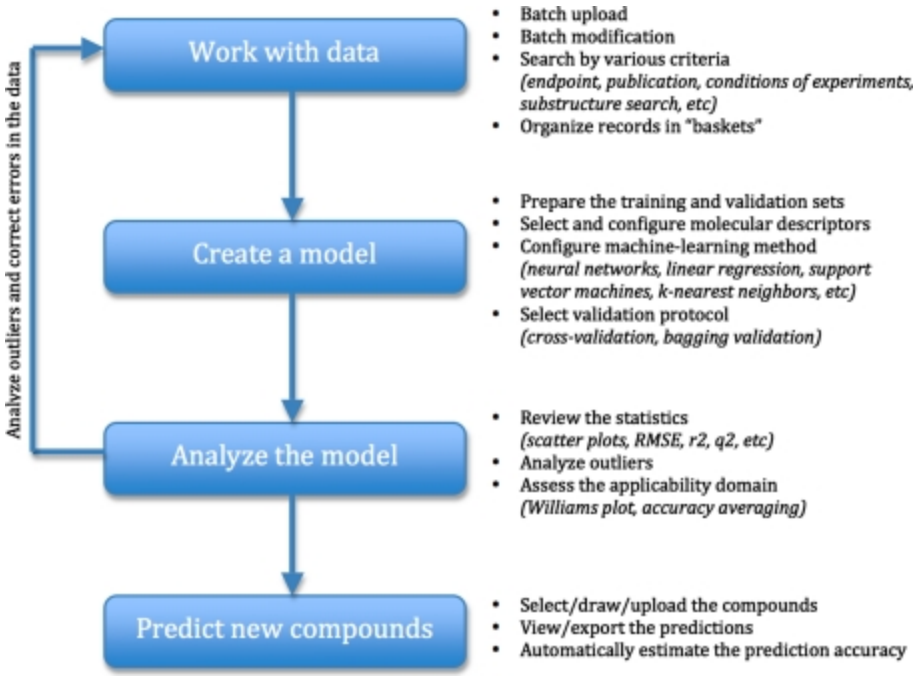
OCHEM modeling framework allows to perform the full cycle of QSAR model development, which includes:

- **Management of datasets** with experimental data.
  Users can create and manage reusable datasets referred to as *baskets*.
- Calculation of **molecular descriptors**.
  OCHEM supports more that 20 types of state-of-the-art molecular descriptors from different 3rd party vendors.
- Running a **machine learning method**
- Proper **validation** protocol of the model
- Calculation of model **statistics**
- **Application** of the model to new compounds
- **Recalculation** of the model based on new experimental evidence

Concisely, the main features of the modeling framework within the OCHEM include:

- Support of regression and classification models
- Calculation of various molecular descriptors ranging from molecular fragments to quantum chemical descriptors. Both whole-molecule and per-atom descriptors are supported.
- Tracking of each compound from the training and validation sets
- Basic and detailed model statistics and evaluation of model performance on training and validation sets
- Assessment of applicability domain of the models and their prediction accuracy
- Pre-filtering of descriptors: manual (external) selection, de-correlation filter, Unsupervised Forward Selection (UFS)
- Various machine learning methods including both linear and non-linear approaches
- N-fold cross-validation and bagging validation of models
- Multi-learning: models can predict several properties simultaneously
- Combining data with different conditions of measurements and the data in different measurement units
- Distribution of calculations to an internal cluster of Linux and Mac computers
- Scalability and expendability for new descriptors and machine learning methods

The steps of a typical QSAR research in the OCHEM system and the corresponding features are summarized in a diagram in the following figure:


Work with data
Create a model
Analyze the model
Predict new compounds
Analyze outliers and correct errors in the data
Batch upload
Batch modification
Search by various criteria
(endpoint, publication, conditions of experiments, substructure search, etc)
Organize records in "baskets"
Prepare the training and validation sets
Select and configure molecular descriptors
Configure machine-learning method
(neural networks, linear regression, support vector machines, k-nearest neighbors, etc)
Select validation protocol
(cross-validation, bagging validation)
Review the statistics
(scatter plots, RMSE, r2, q2, etc)
Analyze outliers
Assess the applicability domain
(Williams plot, accuracy averaging)
Select/draw/upload the compounds
View/export the predictions
Automatically estimate the prediction accuracy