Glossary

`Bemis-Murcko scaffold`

A Bemis-Murcko scaffold is simply the molecular core of a compound with the side chains removed per a set of predefined rules and with certain hetero atoms being replaced by more generic hetero atom placeholders.

`BoxCox transformation`

A BoxCox transformation belongs to the family of power transformations and it transforms non-normally distributed data to (near)normally distributed data. It only works on positive data points, so your data is shifted if you have negative values in your target data.

`Counterfactual`

A counterfactual molecule is a molecule of high structural similarity with low property similarity relative to a base molecule. As an example, if your base molecule is a strong inhibitor of its target protein, then a counterfactual molecule will be a poor inhibitor that is structurally very similar to the base molecule.

`Featurization`

Molecules have to be translated into vector or matrix format before you can train a machine learning model. This is called featurization of the molecules.

`Hyperparameters`

While training models, an algorithm will optimize its parameters, e.g. the slope and intercept for a linear regression, but the algorithm's hyperparameters are set before training and each algorithm has its own et of hyperparameters that adjusts how it models the data. For example, for a random forest algorithm, hyperparameters include number of trees, leaf size, splitting criteria to name a few.

`Kurtosis`

Kurtosis refers to a distribution being heavy- or light-tailed compared to a normal distribution or, in other words, being a flat or pointy distribution.

`Sanitization`

Sanitization is the processing of SMILES strings to viable molecules.It includes checking the valence state of all atoms, standardization of tautomers, neutralization of molecules (if possible), and removal of hydrogen atoms unless they have an explicitly set isotope, are attached to a chiral atom, or attached to an atom with unusual valence state. If salts or solvents are present in your dataset, the salts will be removed by the sanitizer, and the . If a molecule fails sanitization it is removed from the dataset.

`Skew`

Skew refers to the symmetry of a distribution around the mean.

`Surrogate model`

A surrogate model is an interpretable model that is modelled in the local area around a data point of interest, e.g. the best performin molecule.

`Tanimoto similarity`

Tanimoto similarity ranges between 0 and 1, with 1 being identical molecules, and is calculated as the normalized overlap between the feature representations of two molecules. In this implementation, the selected featurizer is 2D circular fingerprints.

`Target parameter`

The target parameter is the physical parameter that you are training a model to predict. It could be e.g. membrane permeability or binding affinity.

`Thoroughness`

The thoroughness parameter influences the tuning effort during model training. More effort will find better models, but also takes more time. Consider using Preliminary for a first, quick assessment.

`UMAP`

A UMAP is a reduced dimension representation of the chemical complexity. If two molecules are very different according to their circular fingerprints, then they are distant in the UMAP and vice versa. **Note, the exact distance between two clusters is not interpretable and the size of a cluster relative to other cluster sizes is also not interpretable.**

`Yeo-Johnson transformation`

> A Yeo-Johnson transformation belongs to the family of power transformations and it transforms non-normally distributed data to (near)normally distributed data. It accepts both negative and positive values.