# Zelfstudie 22 Mei – literatuurstudie

RDKit 🡪 used to convert molecule to SMILES and SMILES to molecules

Read in files with SMILES:

A screenshot of a computer program

Description automatically generated

If you want to compute the similarity of molecules based on a fingerprint:

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

[My RDKit Cheatsheet | Towards Cheminformatics (xinhaoli74.github.io)](https://xinhaoli74.github.io/blog/rdkit/2021/01/06/rdkit.html)

Conceptually, molecular representations of different complexity levels (i.e., “dimensionality”) can be used, as follows.

* 0-Dimensional (0D). This representation is independent of any knowledge about molecular structure and atom connectivity. Hence, molecular descriptors obtained from the chemical formula are referred to as 0D descriptors and capture bulk properties (e.g., molecular weight). Some examples of 0D descriptors are atom counts, molecular weight, and sum or average of atomic properties (e.g., sum of atomic van der Waals volumes, mean atomic polarizability).
* 1-Dimensional (1D). Molecules are represented through a list of their substructures, such as molecular fragments, functional groups or substituents of interest. This representation does not require the complete knowledge of molecular structure. The derived descriptors usually are binary (presence/absence of given substructures) or occurrence frequencies.
* 2-Dimensional (2D). This representation considers how the atoms are connected, in terms of presence and nature of chemical bonds. Usually, the molecule is perceived as a graph, whose edges are the bonds and the vertices the atoms. Descriptors based on a graph representation of the molecule encode topological properties (e.g., adjacency, connectivity) and are usually sensitive to structural features such as size, shape, symmetry, branching, and cyclicity. Often, specific chemical properties of atoms are considered, e.g., mass and polarizability, or presence of hydrogen bond donors/acceptors.
* 3-Dimensional (3D). The 3D representation views a molecule as a geometrical object in space and, in addition to the nature and connectivity of the atoms, it accounts for their spatial configuration. In particular, the molecule is defined in terms of atom types and their x-y-z Cartesian coordinates. **Descriptors derived from a 3D representation have a high information content and can be particularly useful for modeling pharmaceutical and biological properties.** *However*, the geometric optimization of molecules can *lead to several issues*, related to: (1) the *influence of the chosen optimization method on the final coordinate values*; (2) the *presence of several similar minimum energy conformers for highly flexible molecules*; and (3) the *difference between the bioactive geometry and the optimized geometry,* the degree of deformation depending upon the number of freely rotatable bonds in the molecule. For these reasons, the **cost/benefit of using 3D descriptors should be evaluated on a case-by-case basis**.
* 4-Dimensional (4D). In addition to the molecular geometry, a “fourth dimension” can be introduced, usually aiming to **quantitatively identify and characterize the interactions between a molecule and a receptor’s active site**.

In particular, molecular descriptors can be divided into classical molecular descriptors and binary fingerprints, as follows:

* “Classical” molecular descriptors (MDs) are designed to **encode a precise structural/chemical feature (or a set of features of different complexity) into one, single number.** Thus, each descriptor can be used alone or in combination with other descriptors. Classical descriptors can have different measurement scales: they can be integers (e.g., number of double bonds and counts of atom types), binary (e.g., presence/absence of a given substituent) or can have continuous values (e.g., molecular weight).
* Binary fingerprints (FPs) give a **complete representation of all of the structural fragments of a molecule in a binary form**. Unlike classical descriptors, fingerprints encode information about 1D-2D molecular structure in a series of binary digits (bits) that represent the presence or absence of particular substructures in the molecule and **are meaningful only when used as a whole**. Typically, a set of patterns (e.g., branched/linear fragments or substructures) are generated from a given molecule, and the presence/absence of a pattern is encoded within a string of a given length and marked as “1” or “0”, respectively. Hashing algorithms are often applied, leading to a “collision” of multiple features in the same bit(s) and to a loss of one-to-one correspondence with molecular features (determining the difference with classical molecular descriptors). **Fingerprints allow performing quick calculations for molecule similarity/ diversity problems but lack the possibility of an immediate connection with precise structural features.**

Dimensionality Reduction

This procedure (also known as variable reduction) aims at reducing the number of variables (i.e., molecular descriptors) by eliminating redundant and irrelevant information. Over the years, different approaches have been proposed. The simplest and most commonly applied procedures to molecular descriptors are:

* Removal of missing-valued descriptors. Some descriptors cannot be calculated for all the molecules and this may limit the model applicability to new molecules. Moreover, often, the presence of missing values may introduce some errors in data analysis and modeling algorithms. As different molecular descriptors often encode overlapping and/or similar chemical information, the most reasonable solution is to delete those with missing values. In peculiar cases, one could choose to retain a descriptor of specific interest and remove the molecules with missing values. This is useful, for instance, when the number of dataset molecules is much larger than that of the descriptors and the number of molecules with missing values is small.
* Low variance filter. A way of measuring the variability of a descriptor for a given dataset is to measure its variance; the lower the variance, the lower the information it contains for the analyzed molecules. Thus, eliminating the descriptors with a low variance allows neglecting irrelevant information.
* High correlation filter. The presence of many correlated descriptors may be an issue in modeling procedures, as it could lead to overemphasizing some structural features and reduce the modeling performance. A molecular descriptor with values highly correlated to those of other descriptors is not going to add much new information to the existing pool of input features, and, thus, it can be removed

Computational Modeling

Once the descriptors have been calculated, they can be used as the independent variables to model a given biological or physicochemical property. They have many applications, the two most important being similarity searching and (quantitative) structure–activity modeling ([Q]SAR).

Similarity searching refers to finding a set of **molecules similar to one or more queries**, which are often chosen because of some optimal properties/bioactivities. The main assumption is that, the more similar the molecules are to the queries, the more likely they are to exhibit the same experimental properties.

**QSAR aims to find a mathematical relationship between a set of molecular descriptors and a given bioactivity/property**. Several steps are required for developing QSAR models, such as: (1) *data splitting* into a training set and a test set, the former used for model calibration, while the latter for model evaluation; (2) the choice of the *appropriate modeling technique*, according to the project scopes and the required performance; (3) a *supervised variable selection* to identify the best descriptors to model the property of interest, and increase model stability, performance and interpretability; (4) *model evaluation through dedicated metrics*. For instance, **for quantitative responses** to be predicted, one useful metric is the root mean squared error (RMSE), expressed as follows: A math equation with black lines

Description automatically generated with medium confidence

where n is the number of molecules considered, while yi and byi are the experimental and the calculated/predicted response of the i-th compound, respectively. RMSE represents the mean model error and has the same measuring unit of the modeled response. When the i-th molecule is used to calibrate the model (i.e., fitting), the RMSE is often referred to as RMSEC (root mean squared error in calculation), while, when the molecule was used only for validation purposes, the acronym RMSEP (root mean squared error in prediction) is often used. **For qualitative responses** (i.e., classification), some useful parameters are Sensitivity (Sn), Specificity (Sn), and non-error rate (NER) [81], expressed as follows: A math equations with black text

Description automatically generated with medium confidence

where TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives, respectively. Usually, correctly classified active compounds are considered as true positives and correctly classified inactives are considered as true negatives. The sensitivity (Sn) represents the ability to correctly identify active compounds, while the specificity (Sp) quantifies the ability to correctly identify the inactives. Finally, the non-error rate (NER) is a measure of the global classification performance, the higher the better.

[Impact of Molecular Descriptors on Computational Models | SpringerLink](https://link.springer.com/protocol/10.1007/978-1-4939-8639-2_5)

Random forest is a commonly used algorithm to find the best descriptors for the application and can also be used to then make predictions.

This article proposes a two stage method based on RF. In the first stage they train the RF with information on known drugs and inactive compounds, to help define the best descriptors for classification. In the second stage they use RF as a classifier to determine the goodness of the selection to provide a prediction of a molecules activity.

[Automatic selection of molecular descriptors using random forest: Application to drug discovery - ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S0957417416306819#:~:text=Random%20Forest%20selects%20automatically%20the,features%20depends%20on%20the%20dataset.)

Tanimoto similarity is often used for similarity analysis between molecules. This article uses univariable regression to find the best descriptors. For the machine learning analysis, random forest regression and bagging regression turned out to perform the best.

[Machine Learning-Assisted Prediction of the Biological Activity of Aromatase Inhibitors and Data Mining to Explore Similar Compounds - PMC (nih.gov)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9798507/)

Traditionally chemical similarity is measured by the Tanimoto coefficient (Tc). A molecular fingerprint, which is a high-dimensional bit vector that captures the presence or absence of chemical groups in a molecule, is used by the Tc to calculate the similarity between compounds. Has some limitations.

Recent years, the t-distributed stochastic neigbor embedding (t-SNE) algorithm has been shown to be a powerful tool to visualize complex high-dimensional data sets in diverse experimental settings.

[Drug Discovery Maps, a Machine Learning Model That Visualizes and Predicts Kinome–Inhibitor Interaction Landscapes | Journal of Chemical Information and Modeling (acs.org)](https://pubs.acs.org/doi/10.1021/acs.jcim.8b00640)

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedTop 50 best features selected by the RFE algorithm with FS = 361 (number of features selected by the algorithm) for antimalarial drug predictions:

The ‘Autocorrelation’ module generates atom type autocorrelation descriptor values, and the autocorrelation descriptors are the molecular descriptors encoding both molecular structure and physicochemical properties of a molecule and also numerical properties assigned and attributed to atoms. These descriptors are calculated by Moreau–Broto (ATS), Moran (MATS), and Geary (GATS) algorithms from lag 1 to lag 8 for four different weighting schemes. The descriptors from the aforementioned module describe how a considered property is distributed in the topological molecular structure, and have a crucial influence on the antimalarial activity prediction.

[Antimalarial Drug Predictions Using Molecular Descriptors and Machine Learning against Plasmodium Falciparum - PubMed (nih.gov)](https://pubmed.ncbi.nlm.nih.gov/34944394/)

Random forest regression

Random forest operates by constructing a multitude of [decision trees](https://builtin.com/data-science/classification-tree) at training time and outputting the class that’s the mode of [the classes (classification)](https://builtin.com/data-science/random-forest-python-deep-dive) or [mean prediction (regression) of the individual trees](https://builtin.com/data-science/regression-tree).

A random forest is a meta-estimator (i.e. it combines the result of multiple predictions), which aggregates many decision trees with some helpful modifications:

1. The number of features that can be split at each node is limited to some percentage of the total (which is known as the **hyper-parameter**). This limitation ensures that the ensemble model **does not rely too heavily on any individual feature** and makes **fair use of all potentially predictive features**.
2. Each tree draws a random sample from the original data set when generating its splits, adding a further element of randomness that prevents [**overfitting**](https://builtin.com/data-science/model-fit).

The above modifications help prevent the trees from being too highly correlated.

[Random Forest Regression in Python Explained | Built In](https://builtin.com/data-science/random-forest-python)

Tanimoto similarity

The Tanimoto algorithm provides a measure of similarity between two sets of fingerprint "bits," denoted as A and B. The Tanimoto coefficient, T(A,B), is calculated as the ratio of the intersection of A and B to the union of A and B, represented by **T(A,B) = |A ∩ B| / (|A| + |B| - |A ∩ B|)**. This coefficient ranges from 0, indicating no common bits between the fingerprints, to 1, representing identical fingerprints. Consequently, a chemical similarity problem would involve finding formulas with a Tanimoto coefficient above a specified threshold, where higher thresholds indicate greater similarity between molecules.

[Tanimoto Similarity and Jaccard Indexes with FeatureBase](https://www.featurebase.com/blog/tanimoto-similarity-in-featurebase)

## Summary of most important insights:

* There are two different types of molecular descriptors (classical molecular descriptors and binary fingerprints)
* 2D descriptors are probably best to use for our situation
* We can reduce the amount of variable we take into account for the model using for example ‘removal of missing-valued descriptors’, low variance filter or high correlation filter
* The steps we need to undertake for the development of our QSAR model are:
  + Data splitting into training and test set
  + Choose appropriate modelling technique (random forest or bagging regression for machine learning, chemical similarity using Tanimoto coefficient)
  + Use a supervised variable selection (random forest or univariable regression)
  + Evaluate model through dedicated metrics (like RMSE [quantitative], sensitivity/selectivity/non-error rate [qualitative])