

Featurizers

RDKit Descriptors

A descriptor of 210 bits where each bit represents a physical property such as number of valence electrons or the presence or absence of a small functional group.

MACCS keys

A descriptor of 166 bits where each bit position represents the presence or absence of a pre-defined structural feature such as functional groups. You can read more about the algorithm here: Reoptimization of MDL keys for use in drug discovery, J Chem Inf Model, 2002, 42, 1273

2D circular fingerprint

In circular fingerprints, a molecule is broken down into all possible substructures. The largest allowed size of the substructures are determined by the radius. Our implementation uses a radius of 6, a bit vector of size 2048, and explicitly considers chirality. The featurizer is also known as ECFP (=extended circular fingerprint), and is similar to Morgan fingerprints. You can read more about the algorithm here: Extended-Connectivity Fingerprints, J Chem Inf Model, 2010, 50, 742

2D pharmacophore

The molecules is translated to a 2D map of pharmacophore elements - hydrogen bond acceptor, hydrogen bond donor, basic group, acidic group, hydrophobic group, halogen, attachment point to an aliphatic ring, and attachment point to an aromatic ring. Results in a vector of size 39972. You can read more about the algorithm here: Genetic optimization of combinatorial libraries, Biotechnol Bioeng, 1998, 61, 47

3D circular fingerprint

Analogous to 2D circular fingerprints, but instead of mapping substructures based on connectivity it is mapped based on spacial proximity. Our implementation uses a radius multiplier of 1.5 Å and results in a vector of size 2048. The featurizer is also known as E3FP. You can read more about the algorithm here: A Simple Representation Of Three-Dimensional Molecular Structure, J Med Chem, 2017, 60, 7393

3D pharmacophore

The molecule is translated to a 3D map of pharmacophore elements - hydrogen bond acceptor, hydrogen bond donor, basic group, acidic group, hydrophobic group, halogen, attachment point to an aliphatic ring, and attachment point to an aromatic ring. Results in a vector of size 39972. You can read more about the algorithm here: Genetic optimization of combinatorial libraries, *Biotechnol Bioeng*, 1998, 61, 47

Coulomb matrix

A representation of the electronic structure of a molecule. Each element in the matrix denotes the electronic interaction strength between two atoms in the molecule. You can read more about the algorithm here: Learning Invariant Representations of Molecules for Atomization Energy Prediction, *NIPS*, 2012, 25

Mordred fingerprint

A descriptor of 1827 bits that are based on a variety of atomic and molecular properties such as molecular weight, bond count, polarizability, and moment of inertia. Our implementation requires a 3D structure of the molecule. You can read more about the algorithm here: Mordred: a molecular descriptor calculator, *J Cheminf*, 2018, 10