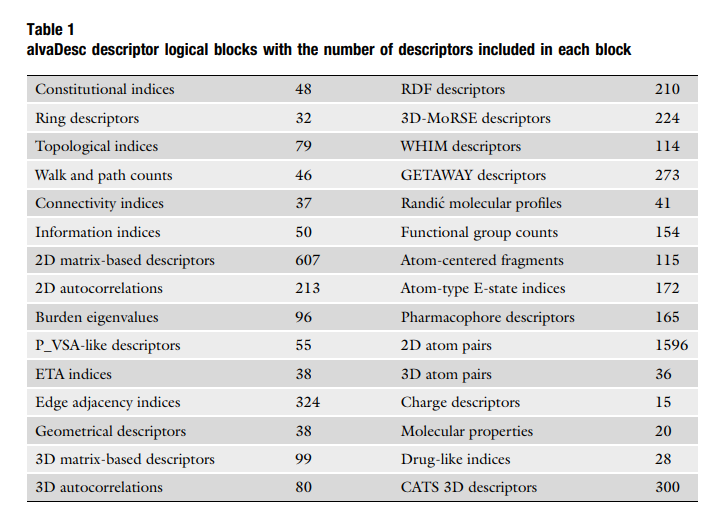
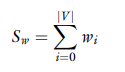
# alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints

* In this the researchers present an alvaDesc, which is a software to calculate and analyze molecular descriptors and fingerprints.
* Molecular descriptors and fingerprints are essential for quantitative structure-activity relationships (QSAR).
* Another incentive for the use of QSAR models is the introduction of regulations on the registration, evaluation, authorization, and restriction of chemicals.
* QSAR models have been proposed for the determination of ecotoxicological properties of *Daphnia magna*, fish, and algae, as well as for chemical prioritization.
* Many libraries, toolkits, and software are available for the calculation of molecular descriptors, among them Mordred [18], Padel-Descriptor [19], CDK [20,21], RDKit [22], and Dragon [23].
* The alvaDesc software calculate the molecular descriptors and fingerprints; it includes 5471 descriptors, 5305 of them are divided into 30 logical blocks.
* The alvaDesc calculate three different molecular fingerprints: -
  + MACCS166 fingerprints [25]
  + Extended-connectivity fingerprints [26]
  + Path fingerprints
* Extended-connectivity fingerprints and path fingerprints can be tuned, not only with respect to the fingerprint size, fragment type, and dimensions, but even by defining atom and bond parameters considered during fragment identifications.



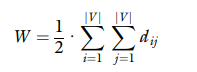
* The alvaDesc provides different tools that carry a first exploration of chemical datasets:
  + Molecule structure verification using PubChem services
  + Molecule structure visualization, charting, and filtering
  + Principal component analysis (PCA) and correlation analysis
* Molecular Structure Curation and Standardization: - Even a single fragment, can be represented in different ways therefore the representation can be different, molecular descriptor values can be affected even if the chemical information is identical.
* Molecular descriptors and fingerprints calculation is based on the assumption that the molecular structure on which the mathematical algorithms are applied to is correct, making molecular structure curation and standardization a fundamental step.
* The alvaDesc software performs its own standardization procedure which includes the nitro-group standardization, the addition of the implicit hydrogens, and the aromaticity detection.
* Molecular Descriptors: - Molecular descriptors are the results of the application of mathematical functions on a well-defined representation of the chemical graph.
* The basic rules of molecular descriptor should comply with: -
  + Be invariant to atom labelling and numbering
  + Be invariant to the molecule roto-translation
  + Be defined by an unambiguous algorithm
  + Have a well-defined applicability on molecular structures
* 0- to 3-dimensional descriptors are the most common chemical structure for information collection.
* The alvaDesc software consist 0- to 3-dimensional descriptors.
* 0-dimensional - it does not contain any information about atom connections, e.g., molecular weight and atom type count.
* 1-dimensional – it consists a part but not all the topology of the chemical structure. e.g., functional group counts, atom-centered fragments, and structural keys.
* 2-dimensional – it includes the information about atomic composition and connectivity of atoms in the molecule. It come from the 2D representation of the chemical structure as a graph.
* 3-dimensional – It consist the use of 3D representation of the molecular graph, the connection between atoms with their position in the 3-dimensional space. e.g., WHIMs (Weighted Holistic Invariant Molecular descriptors).
* Many 0D descriptors are also called circuit rank. They are grouped in alvaDesc in the “Constitutional indices” block.
* The generic formula for this is following: -



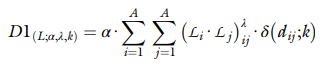
* In the above equation |*V*| is the number of atoms included in the molecular structure and the considered atomic property.
* Molecular weight is a specific case of molecular descriptors as sum of atomic properties where is the atomic mass.
* The cyclomatic number is defined as:



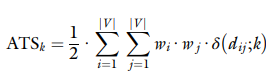
* In this the |*E*| is the number of bonds, |*V*| is the number of atoms, and *D* is the number of disconnected fragments in the molecular structure.
* The cyclomatic number *r* is the minimum number of edges to be removed from a molecular graph in order to remove all its cycles, making it into a forest, i.e., an acyclic graph.
* In the 2-dimensional group, basically all the molecular descriptor are calculated as a topological graph. It can be represented as a matrix.
* The adjacency matrix is the simplest matrix representation of a molecular graph, but also other matrices can be used to represent a graph in order to extract different information.
* One of the most common topological descriptors is the Wiener index (*W*). The Wiener index was originally correlated with the boiling point of alkane molecules; it is a topological index defined as the half-sum of the lengths of the shortest paths between all pairs of vertices in the H-depleted chemical graph [36]:



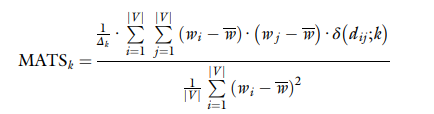
* In the above equation the |*V*| is the number of atoms and is the topological distance between *i*-th and *j*-th atoms.
* The Harary index is used to calculate the reciprocal distance matrix and the Randić connectivity index can be calculated from the matrix.
* Another approach for the definitions of a topological index is the application of mathematical functions to local vertex invariants (LOVIs).
* Autocorrelation is also an example of 2-dimensional descriptors, it can be defined as the following formula:



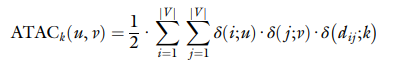
* Here is a generic local vertex invariant; and are a scaling and a power parameter, respectively; and is a Kronecker delta function equal to one for pairs of substructure centers at topological distance = *k* and zero otherwise; *A* is the number of substructure centers that typically are the molecule atoms.
* The Moreau-Broto autocorrelation descriptors (ATS) can be gained from the general equation of *D*1 by simply setting = ½ and =1:



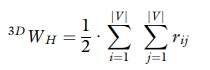
* The Moran autocorrelation descriptors (MATS) can be calculated using the following formula, derived from *D*1 equation:



* In the above equation |*V*| is the set of vertices, and are any atomic property, is the property mean considering the whole molecule, and is the sum of the Kronecker deltas, the number of atom pairs at distance equal to *k*.
* The Moreau-Broto and Moran descriptors, atom-type autocorrelation descriptors are discrete descriptors counting the occurrences of atom pairs at a predefined topological distance *k*.
* The formula for the auto-type autocorrelation descriptors are as follows:



* In the above equation *u* and *v* represent two different atom-types; (*i;u*), (*j;v*) and () are three Kronecker delta functions equal to one if atom *i* is of type *u* and atom *j* is of type *v*; and the topological distance is equal to *k* and zero otherwise.
* Atom-type definition for the calculation of CATS 2D descriptors is based on the concept of “potential pharmacophore points” (PPP), where a PPP is a generalized atom-type defined considering the atom as belonging to one of the following categories:
  + Hydrogen-bond donor (D)
  + Hydrogen-bond acceptor (A)
  + Positive (P)
  + Negative (N)
  + Lipophilic (L)
* It provides 15 possible atom pairs, CATS 2D are calculated on varying the topological distance from 0 to 9 obtaining a vector of 150 frequencies.
* The last molecular descriptors are derived from 3-dimenstional information. An example of a 3-dimensional descriptor is the 3D Wiener index () which is a topological index calculated by analogy with the Wiener index from the geometrical distance matrix as:



* In the above equation is the Euclidean distance between atoms *i* and *j.*
* GETAWAYs (Geometry, Topology, and Atom-Weights AssemblY descriptors) and WHIMs (Weighted Holistic Invariant Molecular descriptors).
* GETAWAY descriptors encode information about the role of each atom determining the whole molecule shape and evaluate the interactions among atoms with respect to their geometrical position.
* H matrix is a common matrix used by GETAWAY, it is symmetric matrix, which is defined as:



* In the above equation **M** is the molecular matrix of the centered Cartesian coordinates (*x, y, z*) for a defined conformation.
* The WHIM descriptors have been proposed in order to collect holistic information about the spatial distribution of molecule atoms, such as information on 3-dimensional molecular size, shape, symmetry, and atomic property distribution.
* WHIM descriptors are based on the calculation of eigenvalues and eigenvectors of a weighted covariance matrix of the centered Cartesian atomic coordinates.
* Molecular Descriptor Analysis and Interpretation: - mainly the descriptors are a mathematical manipulation of different representations of the chemical graph they do not propose to identify a specific chemical or physicochemical feature.
* The most used technique for the analysis of molecular descriptor values is the principal component analysis (PCA).
* A new technique called t-SNE has been proposed recently it tries to keep the low-dimensional representation of very similar data points close together.
* Grisoni proposed an ordinary least squares (OLS) model, it contains seven descriptors: *MlogP2, X0Av, X1Per, SaaaC, VE1\_(B)m, B02[N-O],* and *B03[N-Cl]*.
* *MlogP2* is the squared logarithm of the octanol-water partitioning coefficient (logKOW), as predicted by the Moriguchi model; *B02[N-O]* and *B03[N-Cl]* are atom pair descriptors where *B02[N-O]* is equal to 1 if there is at least 1 pair of *N* and *O* atoms separated by 2 bonds and 0 otherwise; and similarly, *B03[N-Cl]* is equal to 1 if there is at least 1 pair *N-Cl* separated by 3 bonds and 0 otherwise.
* *SaaaC* is an atom-type electro-topological state (E-state) index calculated as the sum of the E-state for all the carbon atoms in the molecule having three aromatic bonds.
* *SaaaC* increases with the number of carbon atoms with three aromatic bonds as well as it increases with their reactivity.
* *XoAv* is the average valence connectivity index of order 0, *X1Per* is a perturbation connectivity index, and *VE1\_B*(*m*) is a 2D matrix-based descriptor obtained from the Burden matrix weighted by mass (*B*(*m*)).
* The behavior of *X0Av* finding is a relation among descriptor values, number of atoms with many valence electrons, and the number of aromatic and unsaturated bonds.
* The *X1Per* variability has been associated with the molecular size and shape and with the presence of heavy heteroatoms and multiple bonds.
* *VE1\_B(m)* variability has been associated with the molecular size and shape and with the presence of heavy heteroatoms and multiple bonds.
* Variable Reduction: - descriptors with constant or missing values can be removed in order to reduce the number of considered variables. It can also be achieved by defining the threshold of the correlation among the considered descriptors.
* One method for unsupervised variable reduction, available in alvaDesc, is called V-WSP algorithm. This algorithm is based on the analysis of the correlation matrix and is a modification of the WSP algorithm for design of experiments.
* Structural Keys and Molecular Fingerprints: - Molecular fingerprints can be of two types, structural keys and hashed molecular fingerprints.
* Structural keys identify the presence or absence of a defined list of structural features, while hashed fingerprints describe the molecule identifying all possible fragments.
* Structural Keys: - structural keys are defined as a set of structural features (e.g., atom-centered fragments, functional groups, pharmacophoric atom-types) that have been prepared in order to discriminate among molecules and eventually to highlight chemical properties of a molecule.
* The alvaDesc also have various structural keys like functional group counts, CATS 2D descriptors etc. It also provides the calculation of the MACCS 166 fingerprints.
* Hashed Chemical Fingerprints: - it is a deterministic function it means that under a predefined set of rules, a specific fragment will always be associated to a defined set of bits in the fingerprint.
* It explores the molecular structure storing all possible identified substructures following a set of rules. The number of substructures identifiable in a molecule set is not predefined, a hashing function is used to reduce a variable-size Boolean vector to a fixed-size one.
* Hashing function has the advantage to transform an indefinite set of structural features to a fixed-length vector but introduce the so-called bit collision; this means that two different fragments may share one or more bits among their bit sets.
* The key feature of the hashed fingerprint is called darkness. It represents the percentage of bits set to one. Higher darkness values lead to false-positive matches, while low values indicates that the fingerprint size can be lowered.
* It does not allow reversible-decoding.
* The hashed fingerprint types are included in alvaDesc:
  + Extended-connectivity fingerprints (ECFP)
  + Path fingerprints (PFP)
* These fingerprints calculation can be customized using a set of parameters:
  + Fingerprint size
  + Number of bits per pattern
  + Minimum fragment length
  + Fragment occurrences
* Fingerprint size is the length of the Boolean vector which affects the darkness. The number of bits per pattern is the number of bits used to encode a substructure.
* Minimum fragment length is the smallest size of the detected substructures. Maximum fragments length is the biggest size of the detected substructures, and it affects the darkness.
* The general fingerprint parameters, atom-type identification can be customized considering the following atom parameters:
  + Atom-type
  + Aromaticity
  + Attached hydrogens
  + Connectivity (total)
  + Total bond order
  + Connectivity (no H)
  + Charge
  + Ring memberships in SSSR
  + Smallest ring size in SSSR
  + Bond order
  + Atom-type
* The selected atom parameters affect the fragments identified during hashed fingerprint calculation.
* Dealing with Disconnected Structures: - with respect to non-full-connected structures, molecular descriptors can be grouped in two.
* The first group includes all the descriptors with a mathematical definition which can be applied to non-full-connected structures preserving their chemical meaning.
* The functional groups, atom-centered fragments, fingerprints, and structural keys are additive descriptors, and their interpretation is identical for full-connected and non-full-connected structures.
* The second group collects those descriptors that have a definition and meaning that cannot be directly extended to non-full-connected structures.
* alvaDesc provides six different theoretical approaches for the calculation of molecular descriptors on such structures:
  + Standard
  + Maximum descriptor value
  + Minimum descriptor value
  + Average descriptor value
  + Sum of descriptor values
  + Retain the biggest fragment
* Maximum descriptor value, minimum descriptor value, retain the biggest fragment approaches consider only one of the disjoint substructures included in the whole molecule.
* Average descriptor value and sum of descriptor values approaches consider every disjoint substructure as an isolated molecule, then the result obtained on the isolated substructures are merged using the average or sum approach.
* Maximum descriptor value approach considers every disjoint substructure as a single molecule. The maximum value obtained on all disjoint substructures is retained.



* In the above equation x is the array including all the disjoint structures in the original molecule.
* Minimum descriptor value approach considers every disjoint substructure as a single molecule. The minimum of the obtained values is retained.



* Here, x is the array including all the disjoint structures in the original molecule.
* Average descriptors value also approach considers every disjoint substructure as a single molecule. The average of the obtained values is retained.



* Here, in the above equation |x| is the cardinality of the vector **x** and *xi* is the descriptor value of the *i*-th substance.
* Sum of descriptor values approach considers every disjoint substructure as a single molecule. The sum of the obtained values is retained.



* Here, *xi* is the descriptor value of the *i*-th substructure.
* The biggest fragment approach means that the molecular descriptors are calculated only for the biggest fully connected structure included in the original molecule.
* The biggest fragment is defined as the biggest fully connected structure with the highest number of non-hydrogen atoms.
* Conclusion: - QSAR models can be defined using different representation of molecules; Molecular descriptors, structural keys, and hashed fingerprint represent different approaches to codify chemical structures in a mathematical way.
* The structural keys can easily be interpreted due to the fact that they identify a well-defined list of chemical fragments, their application can be limited if considered in isolation, since they provide a limited representation of a chemical structure.
* On the other hand, the hashed molecular fingerprints have been proposed to described the whole chemical structure and can be successfully used for QSAR analysis.