# Mold2, Molecular Descriptors from 2D Structures for Chemoinformatics and Toxicoinformatics

* Research applications in chemoinformatics and toxicoinformatics increasingly use representations of molecules in the form of numerical descriptors that capture the structural characteristics and properties of molecules.
* The Mold2 software was developed to enable the rapid calculation of a large and diverse set of descriptors encoding two-dimensional chemical structure information.
* The low computing cost for Mold2 makes it suitable not only for small data set, such as in QSAR, but also for large database cost for Mold2 makes it suitable not only for small data sets, such as in QSAR, but also for large databases in virtual screening.
* Introduction: - the structures themselves are difficult for use in the models. Molecular descriptors are used to extract the structural information in the form of numerical or digital representation that is suitable for model development, serving as the bridge between the molecular structures and physicochemical properties or biological activities of chemicals.
* In 1937, Louis L. Hammett developed an equation that become a milestone in chemistry, in this equation, the rate or equilibrium constants of side-chain reactions of aromatic acids, phenols, and anilines as well as other compounds are calculated from two parameters: characterizing the nature of the reaction and quantifying the electronic effect of replacing a hydrogen atom by a given substituent in the *meta* or *para* position. The mechanistically based Hammett equation thus became the first QSPR equation in chemistry.
* In 1960 Hansch and Fujia introduced a steric parameter , the hydrophobicity parameter based on the *n*-octanol/water partition coefficients enabled the extension of QSPR to the realm of biology (QSAR) and has since been extensively used to develop quantitative models from *in vitro* activity data.
* In the most general sense, an effective molecular descriptor is the one representing chemical structure features or chemical properties that vary across a set of chemicals similarly to a biological end point associated with the chemicals.
* Regression and classification models derived from very large descriptor sets of esoteric physical meaning often yielded models with greater accuracy and fidelity.
* The specter of an overfitted model demanded both diligence in descriptor selection and vigilance in guarding against over-fitted models, both of which add an immense computational expense to model development.
* While highly desirable, the *a priori* design and validation of a relatively smaller number of simpler molecular descriptors that can give a comparable or a better result than a large descriptor set for drug candidate discovery and toxicity prediction is a formidable task.
* Mold2 molecular descriptors are easily and quickly calculated with no missing values, a common problem with most existing commercial systems.
* The researchers show that Shannon entropy analysis that he Mold2 descriptors convey a similar amount of information as the other tested descriptor sets.
* MOLD2 DESCRIPTORS: - 1D molecular descriptors present bulk properties of compounds, such as the number of specific atoms, molecular weight, etc., and can be calculated solely on a molecular formula.
* 2D molecular descriptors present structural information that can be computed from 2D structure of a molecule, such as the number of benzene rings, the number of hydrogen bond donors, etc.
* 3D molecular descriptors present structural information that has to be derived from 3D representation of a molecule, such as solvent accessible surface area with positive partial charge in the structure.
* An optimal set of molecular descriptors in the absolute sense is currently not available and may not exist.
* The 1D descriptors are calculated solely based on the molecular formula. The atom counts include numbers of different atoms and the total number of atoms in the molecule. The physicochemical properties are molecular weight and average molecular weight.
* The 2D descriptors are calculated from the 2D structure of a molecule, though some of them such as lop and fragment counts are called 1D descriptors.
* Types of carbon atoms in the 2D descriptors are distinguished based on hybridization status, such as primary carbon, tertiary carbon on ring structure, unsubstituted aromatic carbon, and so on.
* It is also related to the bond information such as numbers of single bonds, double bonds, aromatic bonds, rotatable bonds, and so on.
* The physicochemical properties are calculated from the 2D structure of a compound, which are not similar to the ones in 1D descriptors that are obtained only based on the molecular formula.
* Structural features are specific structural components whose chemical and biological functionality is not explicitly understood, like number of independent rings, circuits, aromatic rings, three-membered rings etc.
* The 2D autocorrelation descriptors *A*(*d*), are calculated from the 2D structure based on the autocorrelation function, in which *d* is a topological distance and can be any number between 1 and the maximum of distance in a molecule, is a function of the variable , *a* is the number of atoms in the molecule, and and are the properties of atoms *I* and *j*.





* The Balaban index descriptors are obtained from the Balaban distance connectivity index *J* that is calculated using the following formula



* In this and are the vertex distance degrees of two atoms connected by bond *k*, *B* is the number of bonds of the molecule, and *C* is the cyclomatic number.
* Descriptors such as the cyclicity index and average cyclicity index are part of the detour index that is calculated from the detour distance matrix of a molecule. The detour matrix (or maximum path matrix) is a square symmetric matrix



* Its element of row *i* and column *j* () is the maximum number of edges between nodes *i* and *j* and is zero from a node to itself.
* The topological distance index descriptors are derived from the distance matrix **D** of a molecule.



* Its elements represent the shortest paths in the number of edges from an atom to another, and by conversion it is zero from an atom to itself.
* Descriptors extracted from the eigenvalues of an adjacent matrix and other matrices of a molecule are classified into the group of eigenvalue-based descriptors, including the Lovasz-Pelikan index, the folding degree index, the characteristic root index, and so on.
* *Ic* is used to measure the degree of diversity of the atoms or bonds in a molecule and defined by the formula



* Here *C* is the number of different types of atoms or bonds, and *nc* is the number of atoms or bonds of the *c*th type.
* Descriptors related to molecular walk counts are calculated based on the graph walks. They are extracted from the adjacency matrix of a molecule. Total walk count, weighted walk degrees, and walk connectivity indices are examples.
* The subclass of Schultz index *SI* descriptors are calculated based on the formula



* Here *a* is the number of atoms or bonds in a molecule, *M* is the adjacency matrix, *D* is the distance matrix, and *v* is the vertex or edge degree vector.
* Descriptors related to the topological charge index are derived from the adjacency matrix and distance matrix of a molecule.
* The Wiener index was firstly defined only for acyclic graphs. The Wiener index descriptors in Mold2 are derived from the Wiener matrix modified from the distance matrix of a molecule.
* The Zagreb index descriptors are derived from the vertex degree of atoms of a molecule, including the quadratic index and the binormalized quadratic index.
* Generation of Mold2 Descriptors: - Mold2 accepts an SDfile of the molecules for which descriptors are to be calculated. Other formats for representing molecular structure have to be converted to an SDfile format for use by Mold2.
* The chemical structures in an SDfile are processed sequentially.
* The main part of Mold2 is a module for computing the 779 descriptors. A set of generalized functions is used to speed up calculations. Among the most important ones is a module for the perception of the smallest set of smallest rings (SSSR) that is adopted from a previously reported algorithm.
* After, all 779 molecular descriptors for a molecule are calculated, results are output to a file; molecules are serially processed until all molecules in the SDfile are processed.
* Evaluation of Mold2 Descriptors: - evaluation of the performance of molecular descriptors is necessary for guiding their proper use.
* There are three ways of evaluating the utility of descriptors. The first is to assess information that is presented in a data set represented by a set of descriptors.
* In general, higher variance in the descriptors corresponds to a high probability of developing a valid model using the descriptors.
* The second is to make sure that there are not any redundant descriptors and not many highly correlated descriptors.
* The third is with a comparative analysis of sets of descriptors when the same modeling approach is applied to the same data set; the differences in modeling results can then be used to differentiate the efficacy of two sets of descriptors.
* Information Content by Shannon Entropy Analysis. The concept of Shannon entropy, also called information entropy, has played a central role in information theory.
* The entropy of a random variable is associated with its probability distribution, which is formulated as



* Here is the probability of outcome *i.* Golden et al. used Shannon entropy to analyze different descriptors by ranking them according to variance, with variance descriptors preferred for discriminating compounds.
* Shannon entropy analysis was conducted for comparing Mold2 descriptors with three sets. First, descriptors nor having a value across all the chemicals in a data set were discarded.
* Next, descriptors were binned for each descriptor in the data sets and the probability distribution calculated using 20 even bins that spread from the minimum to the maximum values of the descriptor.
* Lastly, the comparison of descriptor sets was done assuming that mean Shannon entropy was proportional to the average information encoded in different sets of descriptors for the same data set.
* The distribution of Shannon entropy is another important property of the descriptors, as it is expected that the higher entropy descriptors are, the better a data set is represented by the descriptors.
* Correlations between Descriptors: - if it is not preferable that there are many redundant or highly correlated descriptors in a set of descriptors. Correlation coefficients (r) between the descriptors in each of the compared descriptor’s sets were calculated for the four data sets.
* The weak correlations (0.25 < r2 < 1) are less than 15% for all the compared descriptors sets in all the data sets. The high correlations (r2 > 0.8) are less than 1%.
* Classification Performance Using Decision Forest. The comparison of different sets of descriptors using Shannon entropy only reveals comparative differences in variance of the structural information encoded in the descriptors.
* A better set of descriptors not only carries sufficient information but also should be biologically and physiochemically relevant.
* The quality of a QSAR/QSPR model is solely dependent on the correlation of chemical structure variables, which can be investigated through comparing classification models developed from different information-bearing descriptors.
* Decision Forest classification method developed and was applied to all data sets except the EPA data set.
* Ten-fold cross-validation was used, where the data set was first randomly divided into ten equal portions, and each portion was then successively excluded from the training set and predicted by the model developed from the remaining nine portions.
* The prediction accuracy of the cross-validation was taken as the average of prediction accuracy results of the 10 models.
* The 10-fold cross-validation was repeated 100 times to achieve a statistically unbiased estimation of predictive accuracy, sensitivity, and specificity.
* Conclusion – Diverse applications include virtual library generation and screening, similarity and diversity analysis, QSAR/QSPR, and predictive toxicology.
* The 779 Mold2 descriptors presented in the research are calculated from both 1D and 2D chemical structure.