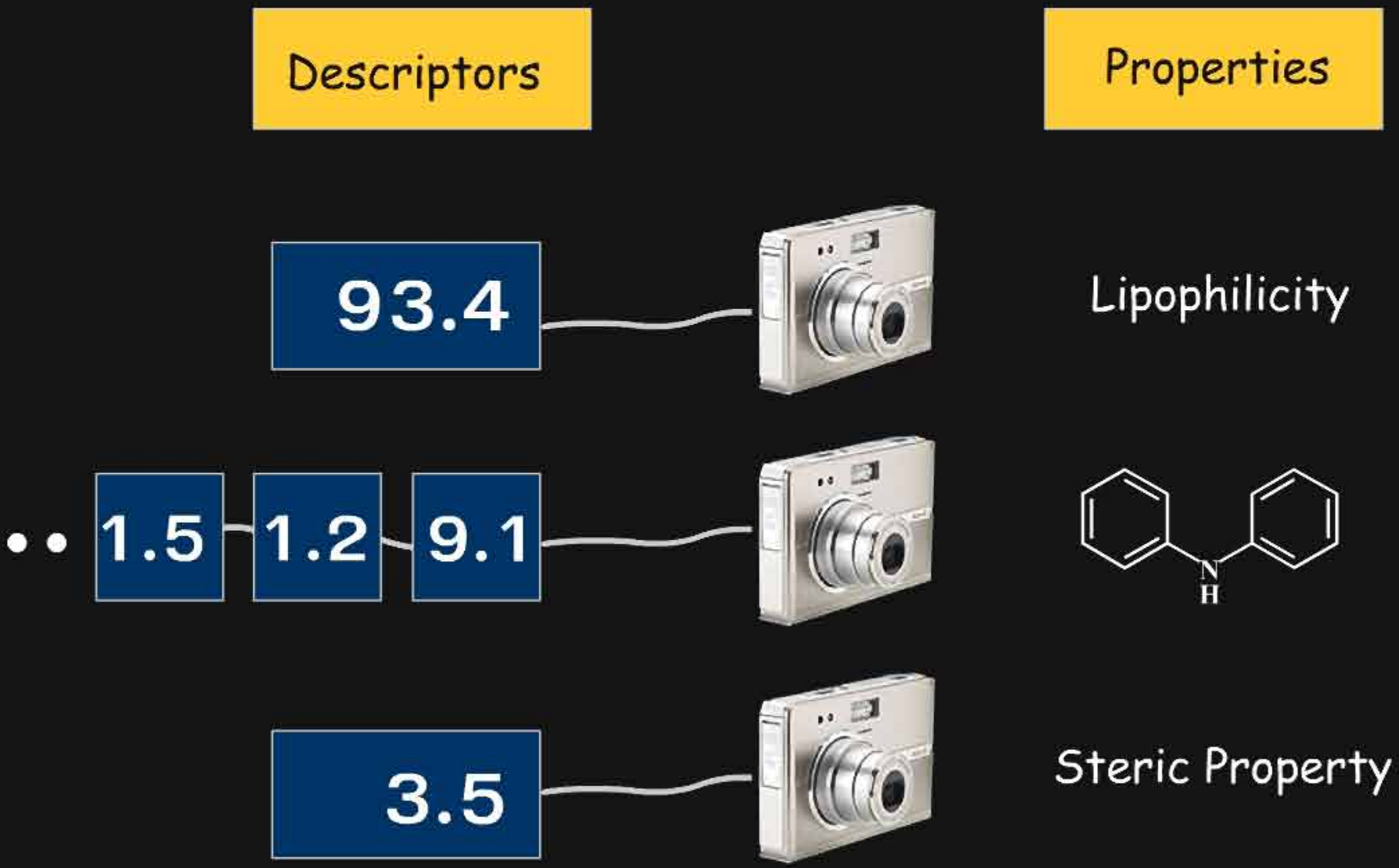
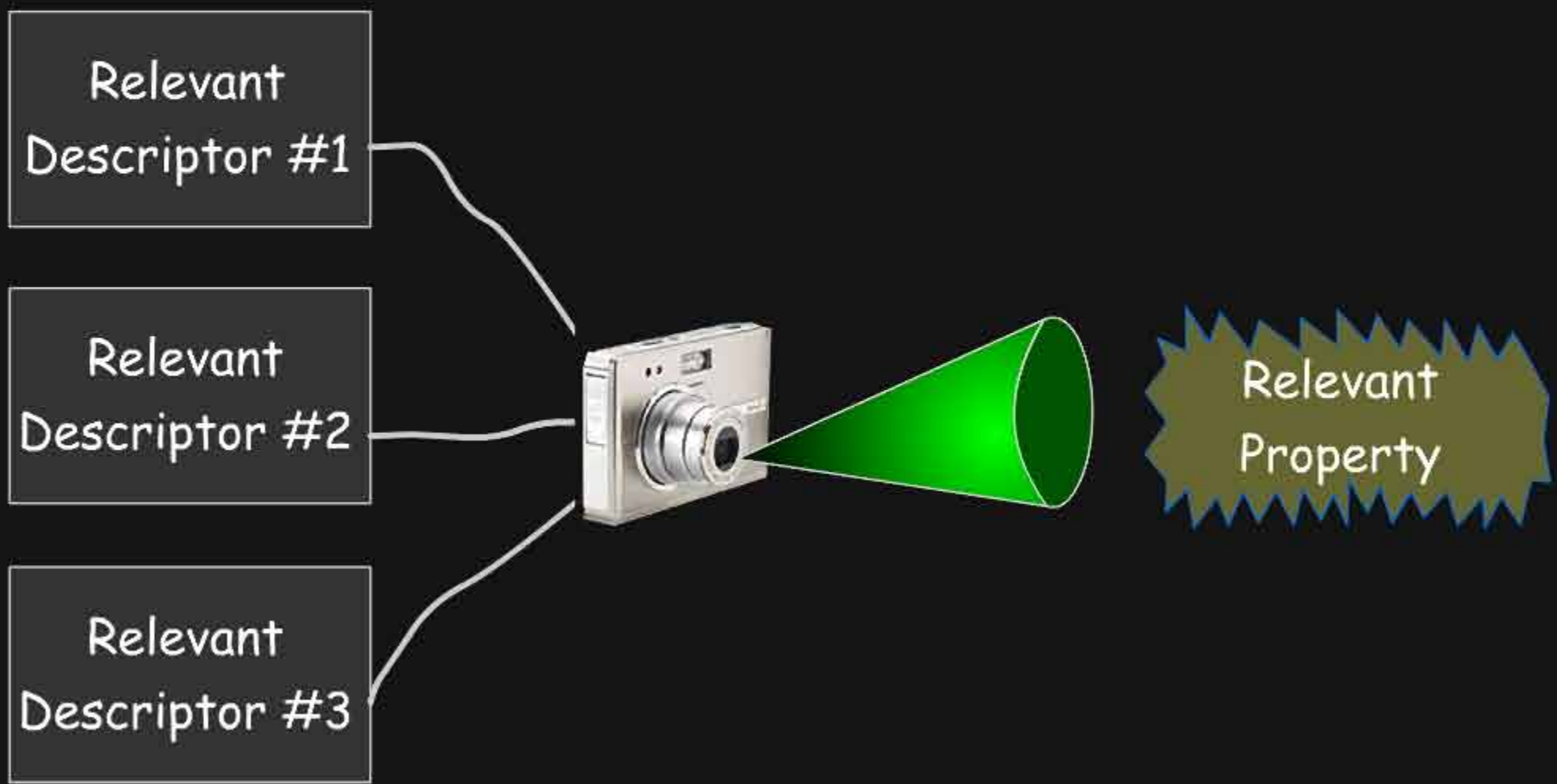
J3.4.1 The Concept of "Molecular Descriptors"

Molecular descriptors consist of numbers that capture the properties and the structures of molecules. They then serve to characterize molecules and compare them by searching through large libraries with respect to physico-chemical, structural and biological properties. Molecular descriptors provide an abstract representation of molecular structures in "descriptor space". This means that structural molecular features are analyzed by an algorithm, and, while trying to capture information relevant to the problem, are translated into a structural format that can be further analyzed by the computer.



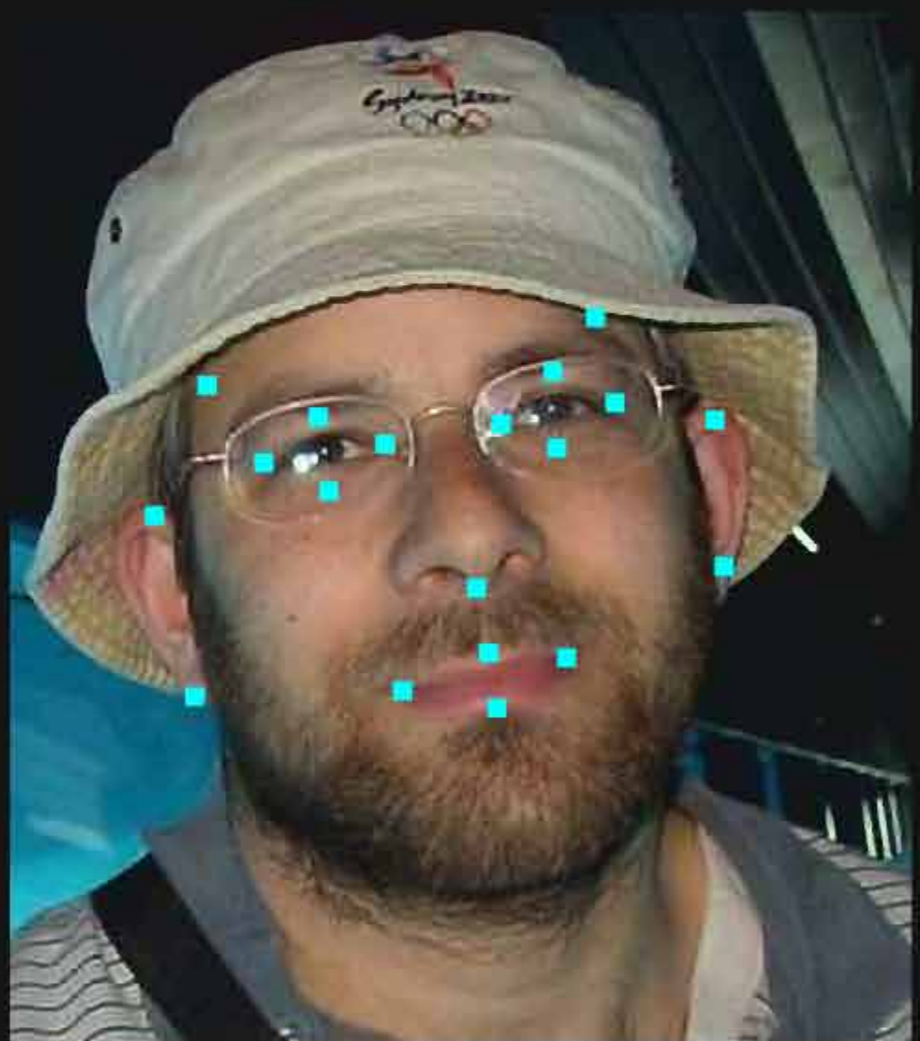
J3.4.3 High-Dimensionality Space of the Molecular Descriptors

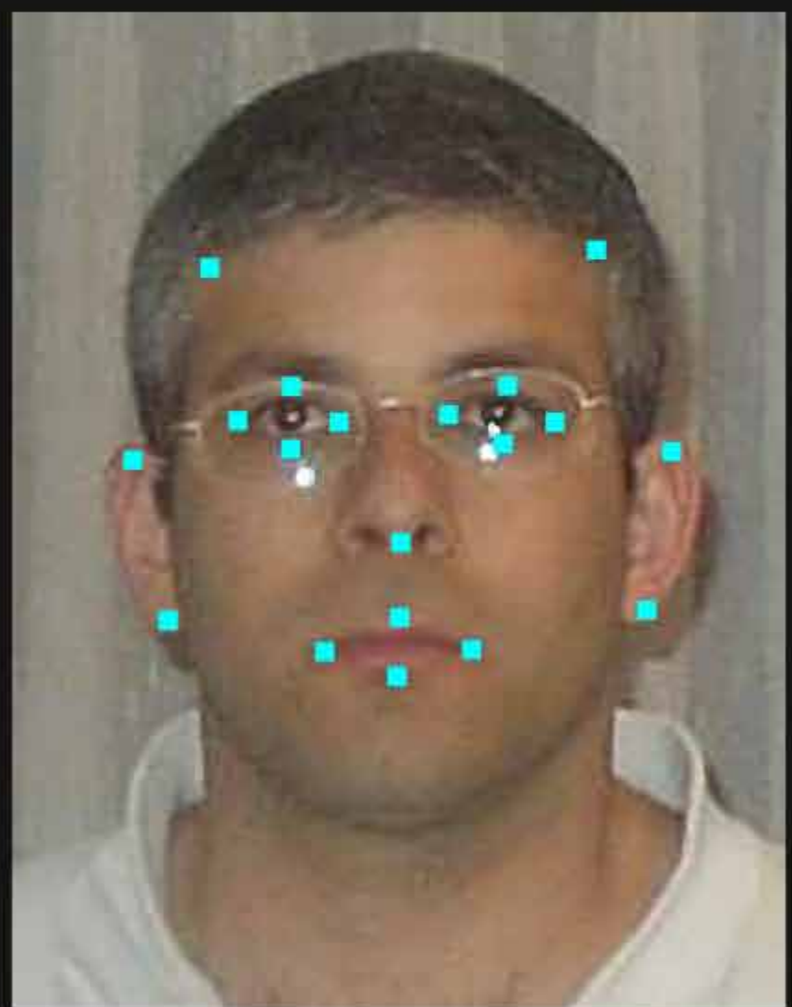
In general many properties must be taken into account to improve the signal-to-noise ratio and several descriptors might be necessary to capture a single property. The "chemical space" will be characterized by a high degree of dimensionality.



J3.4.4 Example of Selection of Relevant Descriptors

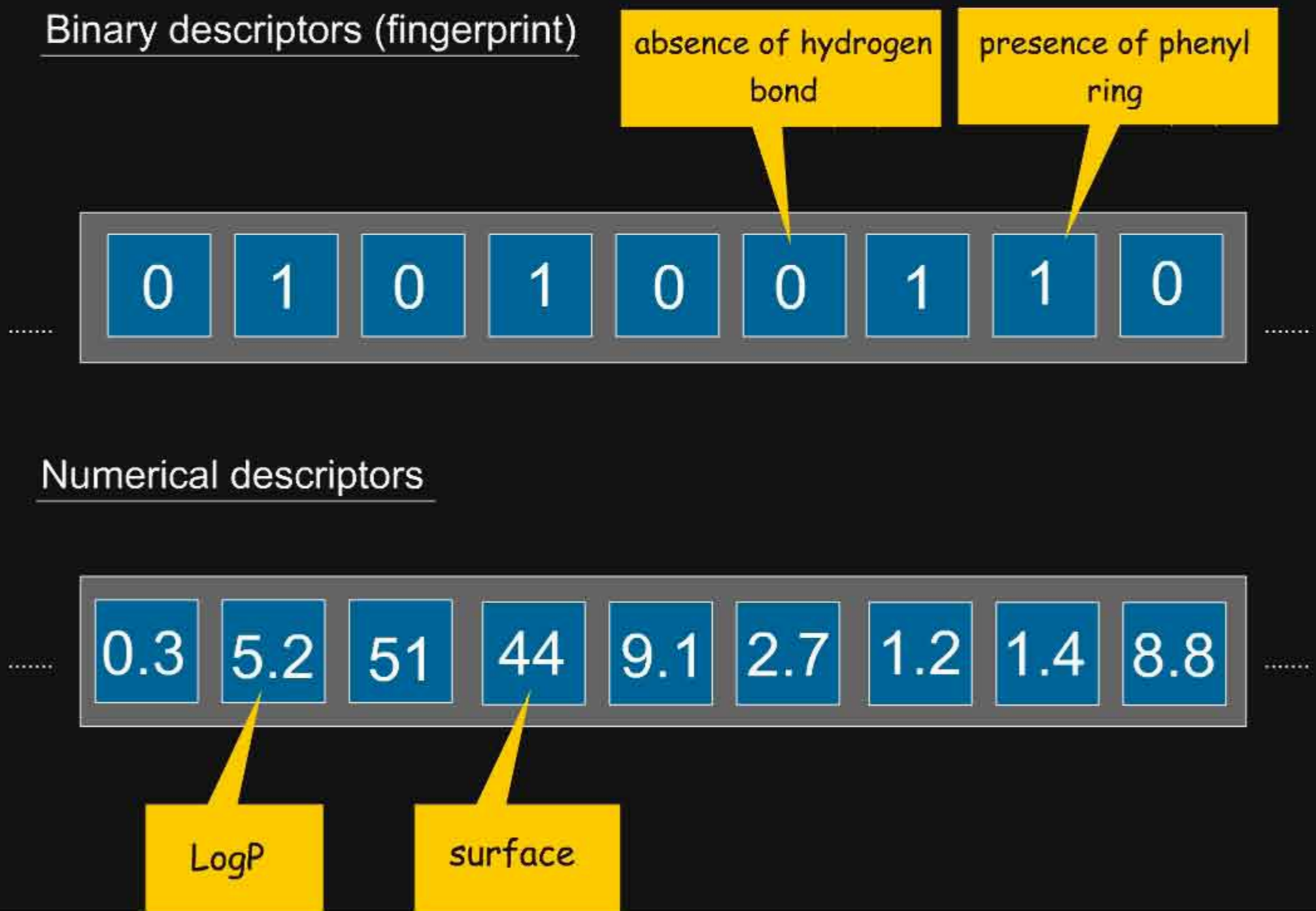
The example below deals with face recognition. The most straightforward way to recognize people is to compare individual pixels of the picture to each other. This procedure would however quickly run into problems if the face moved slightly while the picture was being taken or a change in lighting made the picture slightly lighter or darker. Thus, a computer selected characteristic feature from the faces below and the distances between them do not depend on the precise orientation of the face, and they are also insensitive to changes in illumination. They are relevant descriptors.





J3.4.5 Binary and Numerical Descriptors

Descriptors can be either binary or numerical. Binary descriptors or "fingerprints", are binary bit string representations of on/off (1/0) values, each indicating the presence or absence of certain structural features in the molecule. Each bit of the fingerprint represents a descriptor, and builds a descriptor vector. Storing information in a binary presence/absence format is a computationally very efficient way to handle data. Numerical descriptors encode global or local properties of the molecule considered.



J3.4.6 Experimental and Calculated Molecular Descriptors

Molecular descriptors, of which several thousand exist, can either be derived from the structure of the compound on the basis of an algorithm or they can be based on experimental data. Since experimental data are much harder to obtain than algorithms that can be applied on a structure, most descriptors are algorithmically defined.

17.8	1.2	2.5

0.8	31.2	1.7	0.27	6.8	3.1
52.1	19.7	6.9	0.3	34.2	0.31

experimental descriptors

calculated descriptors

J3.4.7 Predefined vs. Algorithmically Defined Descriptors

Calculated molecular descriptors are either pre-defined or generated on the fly. This is illustrated in the figure below. One well-known class of descriptors, MDL keys, encode 166 pre-defined molecular fragments, some of which are shown below. Unity fingerprints on the other hand determine the type of atoms present between a certain number of bonds and encode whatever type of atoms are present. While MDL keys were designed to show discriminatory power between classes of compounds (and were also further optimized for this purpose), algorithmically defined descriptors are in some cases able to capture more relevant information and encode whatever structural information is given.

Fixed descriptor

Algorithmically defined descriptor

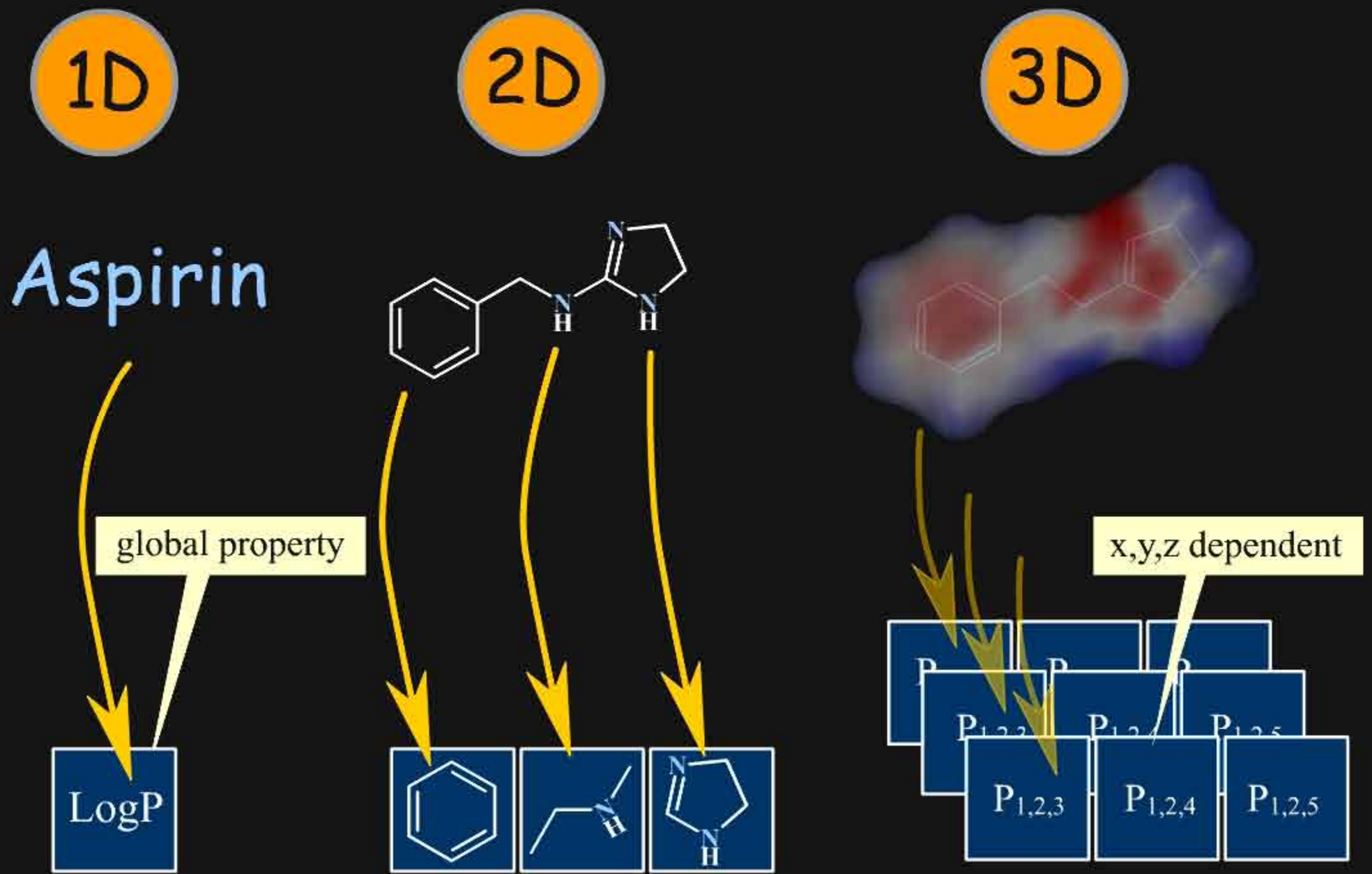
Predefined Fingerprints: E.g. MDL keys (166 bits set):

1 - Isotope
2 - Atomic Number >103 Present
....
84 - NH2 (Primary Amide)
85 - CN(C)C (Cyano Group)
...
165 - Ring Present
166 - Fragments Present

Predefined Fingerprint Sets Make Implicit Assumptions About Relevant Features

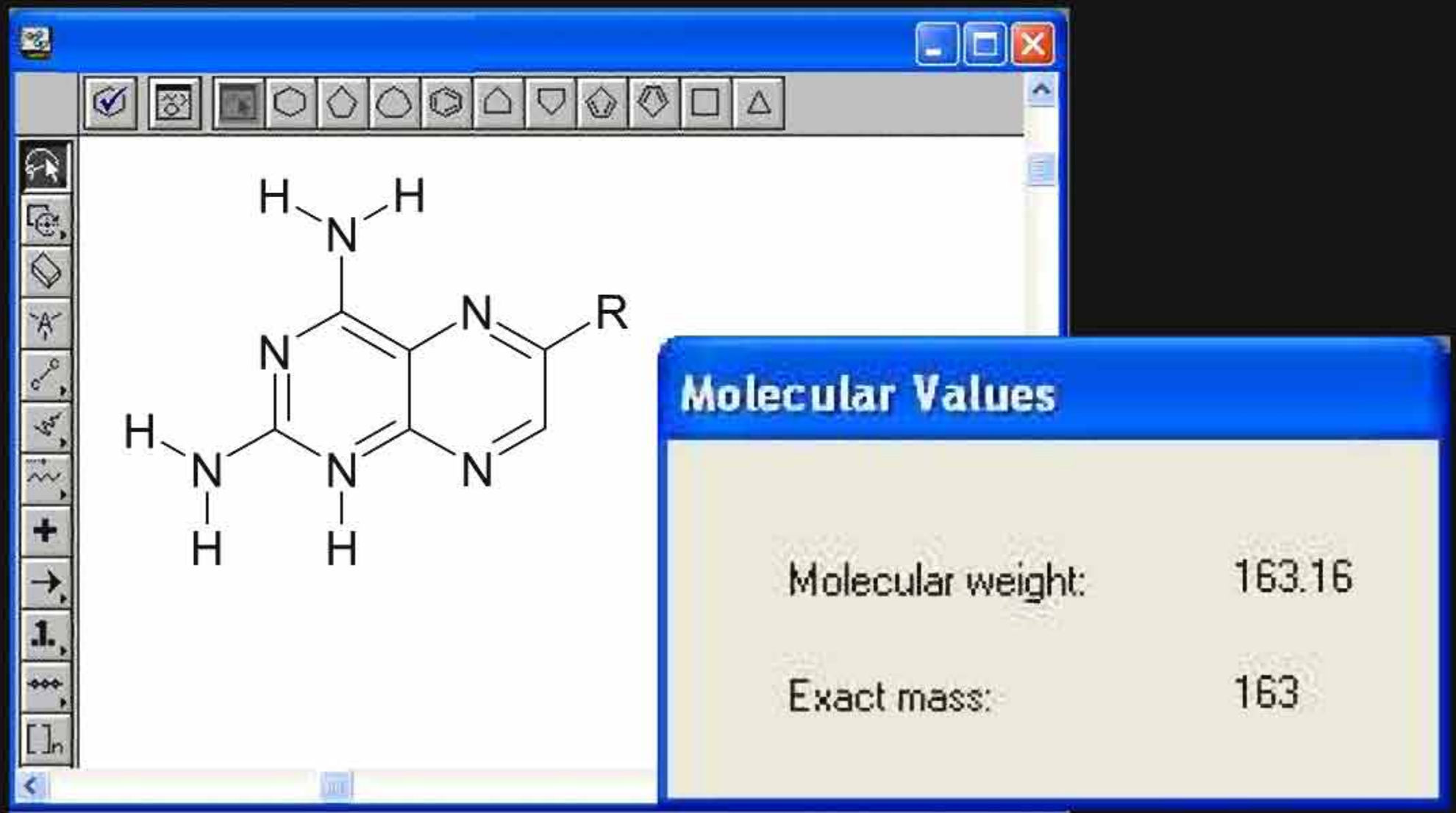
J3.4.8 Possible Classification of Molecular Descriptors

Descriptors are often defined as one-dimensional, two-dimensional and three-dimensional descriptors. This corresponds broadly to descriptors that describe the molecule by a single value (1D) such as molecular weight, or topological indices, descriptors which employ fragments derived from the connectivity table (2D), and descriptors which capture the three-dimensional nature of the molecule (3D). All descriptors can be divided into further subtypes, but the general (although not unambiguous) classification is shown below.



J3.5.1 1D Descriptors: Single Numbers or Sequences

One-dimensional descriptors can be of two different kinds: They can either present the molecule as a single number, for example its molecular weight or its lipophilicity (logP). While this is a very short description of a molecule, it often captures a surprising degree of information relevant for classification such as for example absorption (where molecular weight is a major determinant) or narcotic effect (where lipophilicity is a major determinant of non selectively interrupting cell membranes). One-dimensional molecular representations can also be similar to the sequence of Proteins or DNA, but they are difficult to construct for 'small molecules'.



J3.5.2 Topological Indices

Topological indices are derived from the connectivity table of a molecule and represent information about the topology of the structure, again captured in a single number (the "topological index"). First generation indices such as the Wiener Index are derived from integer values of the connectivity table and are themselves integers. Second generation indices are derived from integer properties but are real-valued numbers while third generation indices employ real-valued numbers for their generation and are also real-valued numbers themselves. Examples are given below. Topological indices have been widely used in QSAR studies, but their interpretation and discriminatory power remains difficult.

Wiener Formula | Method 1 | Method 2

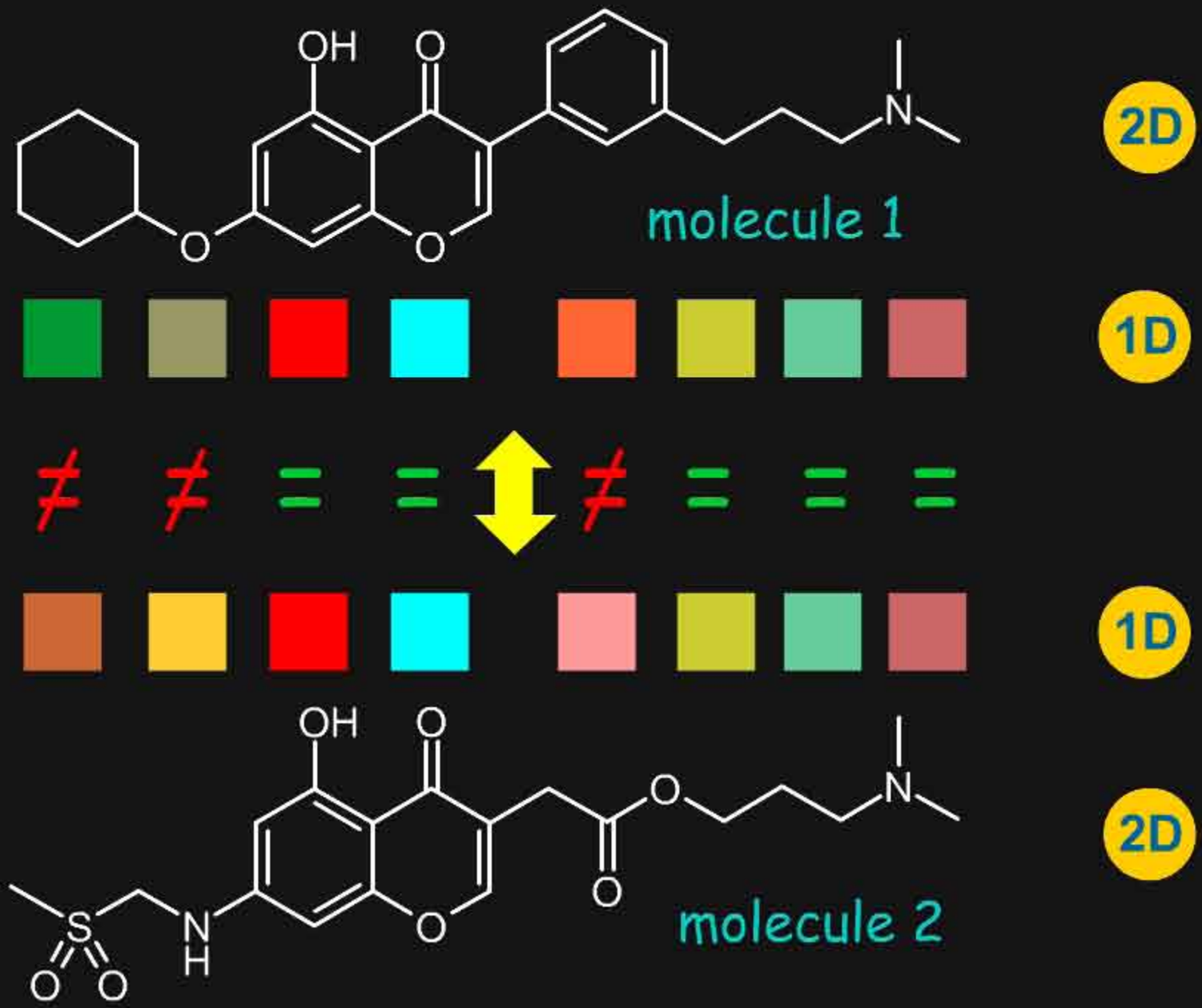
Topological index =

sum of all shortest connections between the atoms along the bonds (by counting the number of bonds)

$$\text{the Wiener index} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} D_{ij}$$

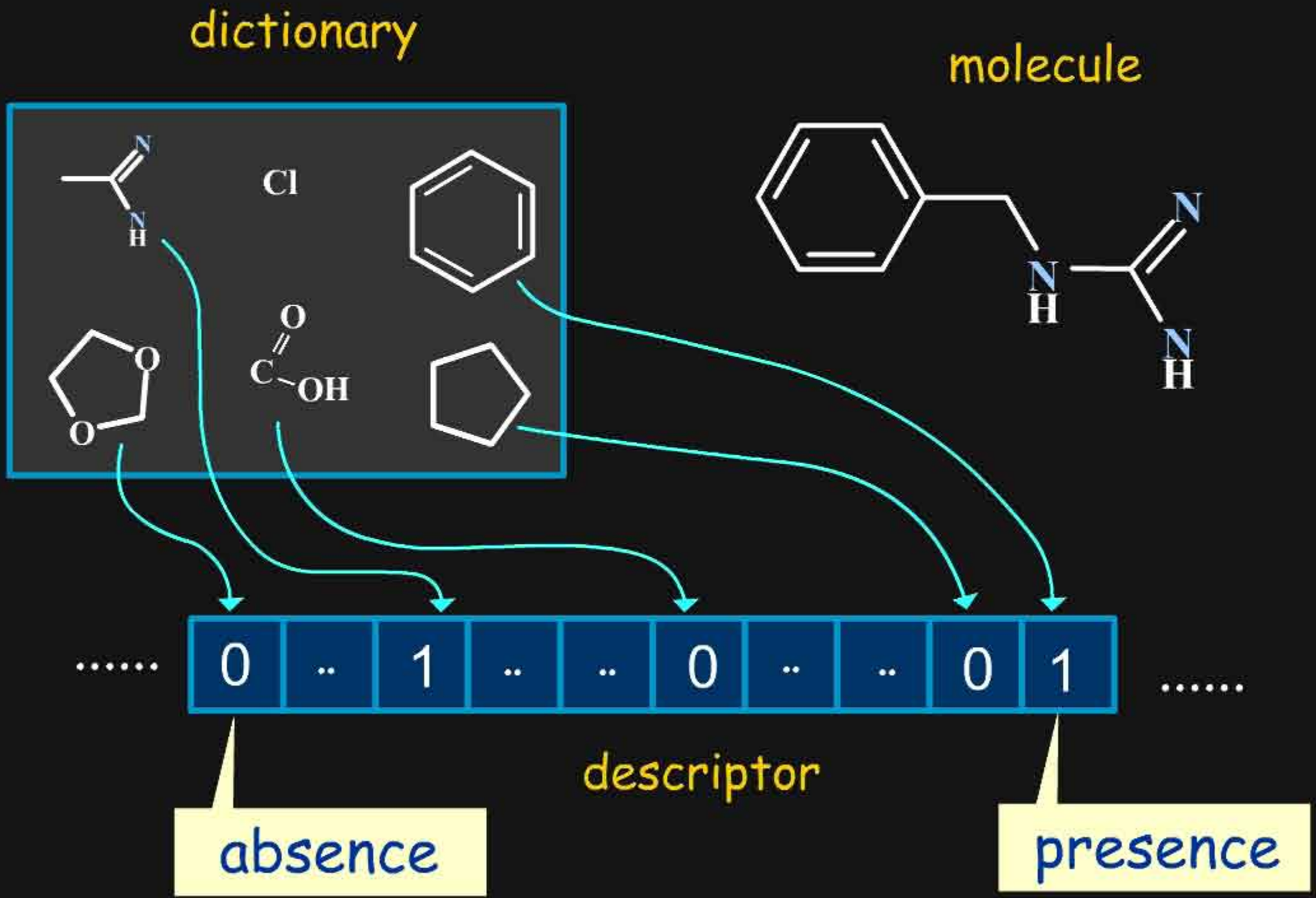
J3.5.4 Linear Representations of Molecules

While DNA and proteins have "natural" one-dimensional representations due to their linear nature, this is not true for "small molecules", which may be and usually are branched and possess cyclic moieties. Nonetheless a method has been devised to represent molecules in a single dimension: this is illustrated below. Molecules are compared by aligning their one-dimensional representations and calculating the overlap of both representations as a similarity measure. Note that while in the example below the method seems to be straightforward, in practice ring systems often introduce complexity.



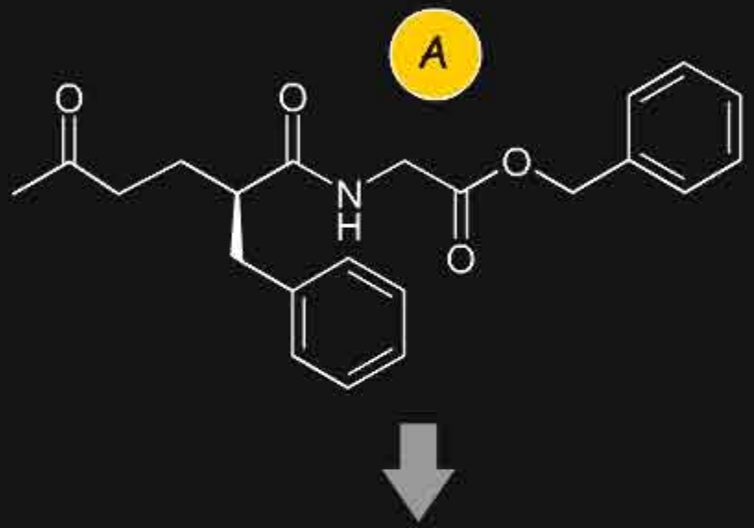
J3.5.5 2D Descriptors: Fragments and Substructures

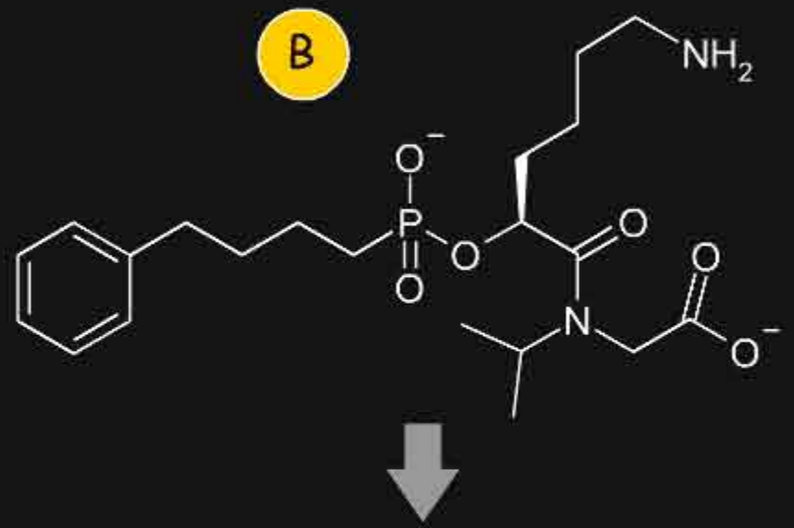
Substructural descriptors "fragment" the molecule into smaller entities and determine whether the resulting fragments match a pre-defined dictionary. They have the advantage of being easy to interpret.

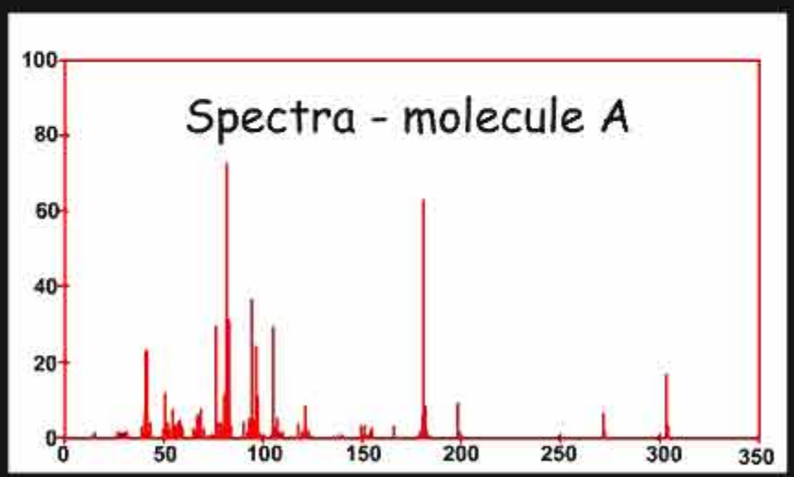


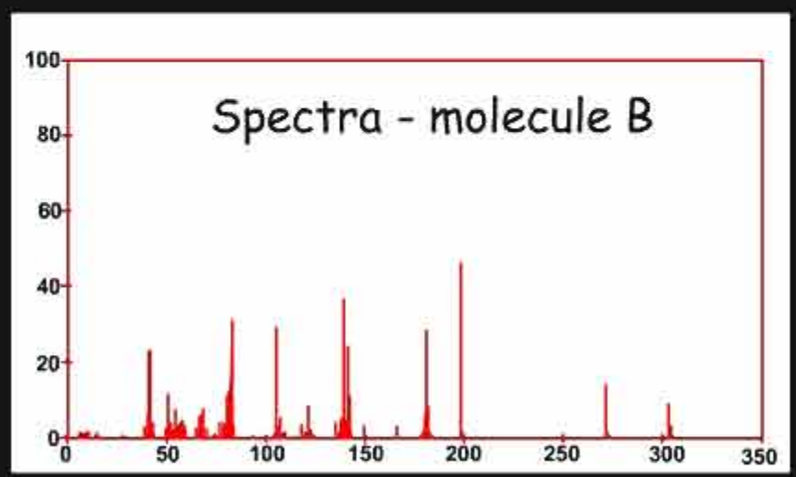
J3.5.6 Spectra-Derived Descriptors

An additional method that does not calculate descriptors from the small molecule directly but uses "natural phenomena" for this purpose is employed by the class of spectra-derived descriptors. Based on either experimental or calculated spectra (which may be IR spectra or others), spectra are encoded, for example by analyzing their zero crossings. One of the methods in this category is the EVA descriptor, which is based on the vibrational absorption spectrum.





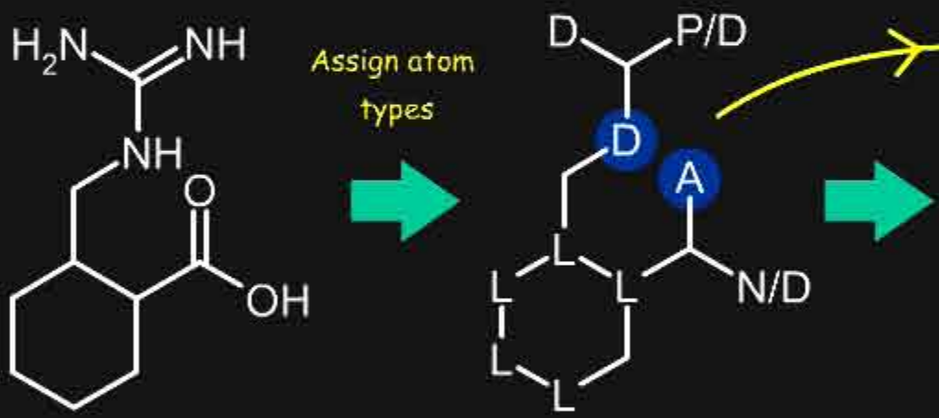




similarity analysis

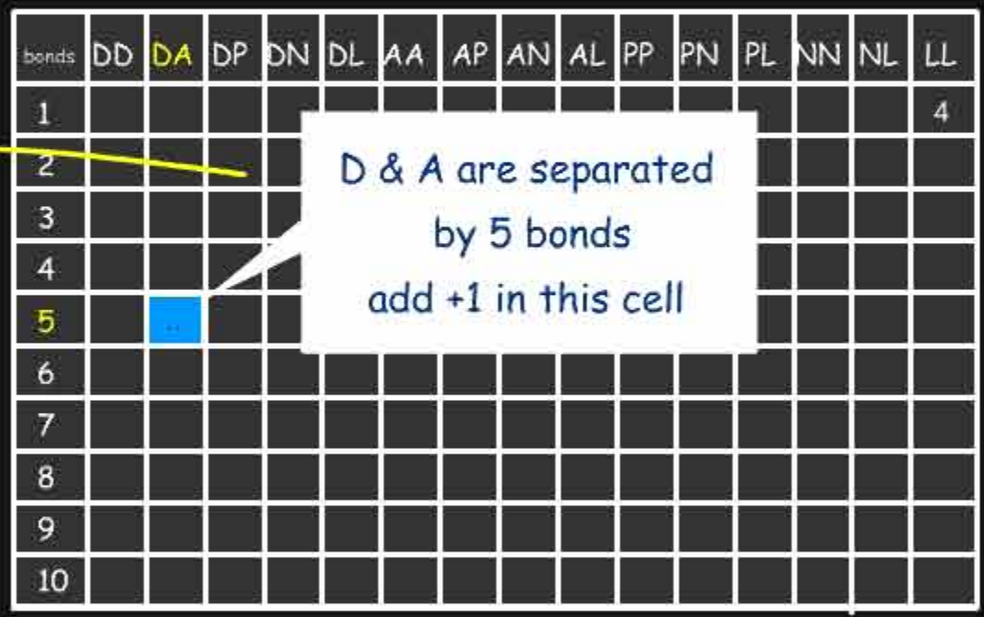
J3.5.7 Graph-Based Multiple Point Pharmacophores

The distance between atoms with putative interaction types can either be defined in space (as in the previous example) or it can be defined as a topological distance (i.e. by counting the number of bonds along the shortest path between two atoms). One example of this descriptor is the CATS descriptor, which employs five interaction types: lipophilic atoms, hydrogen bond donors and acceptors, and positively and negatively charged centers.



Assign atom types

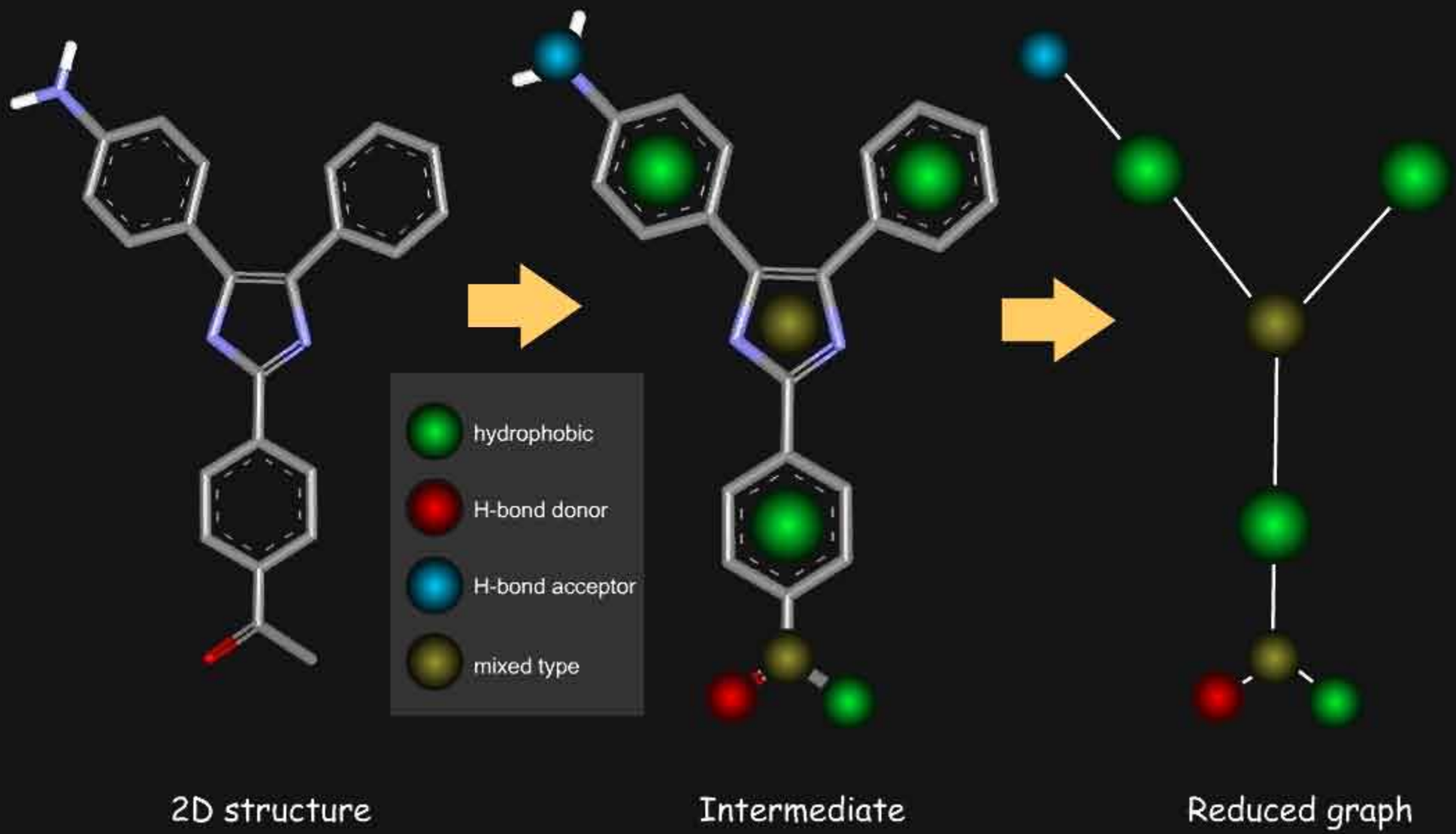
(D) H-bond donnor
(A) H-bond acceptor
(P) positively charged
(N) negatively charged
(L) lipophilic



distribution of pairs of atom types and their separation (number of bonds)

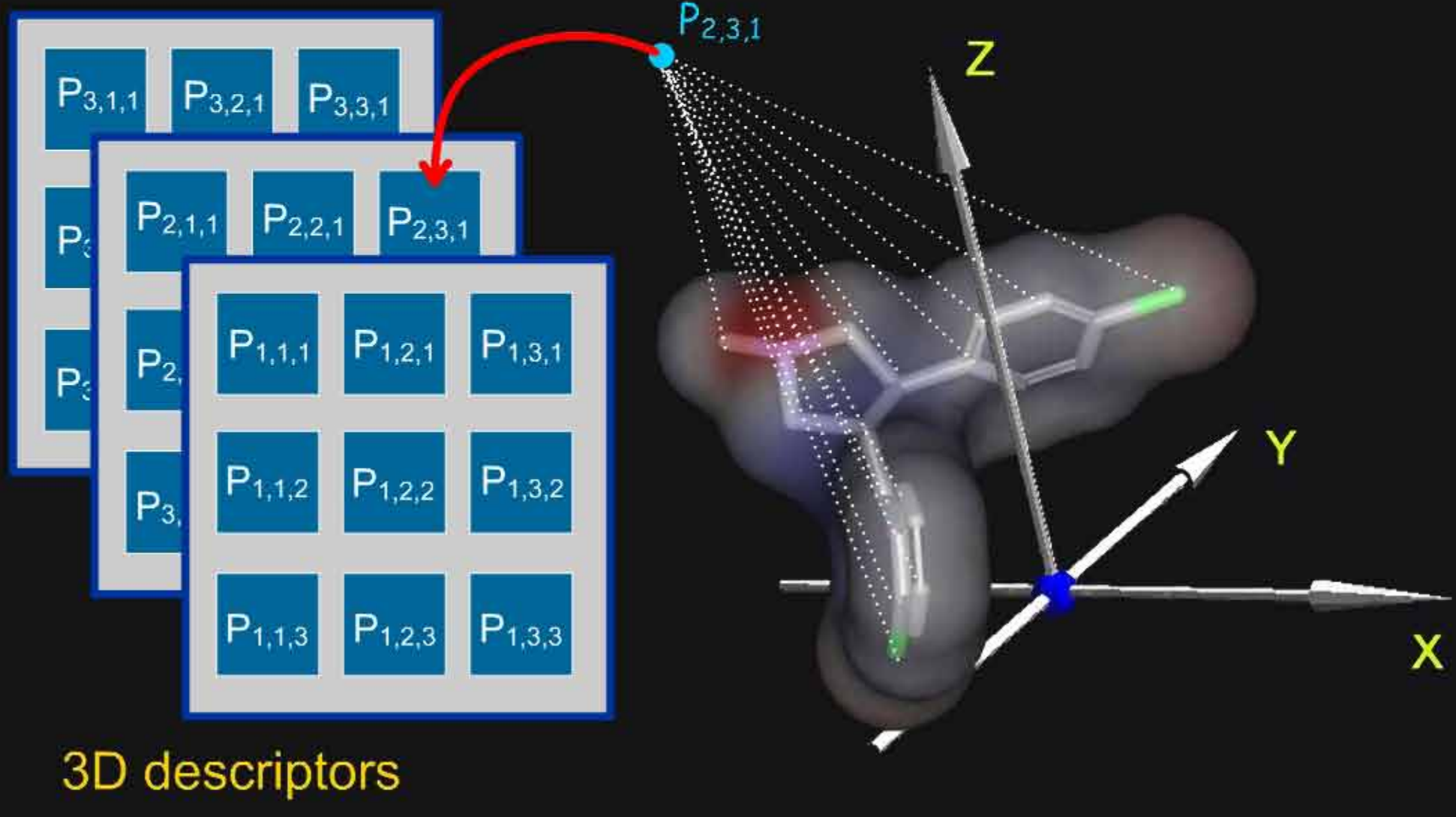
J3.5.8 Reduced Graph and Feature Trees

Reduced graph and feature tree representations of molecules are similar in that they represent the molecule as a graph of edges and nodes, where nodes represent atoms (or groups of atoms) and edges represent bonds (or connections between groups of atoms). Feature trees compare two molecules by partial matching of their graph representations and the nodes include additional information about the atoms such as volume and electronic information.



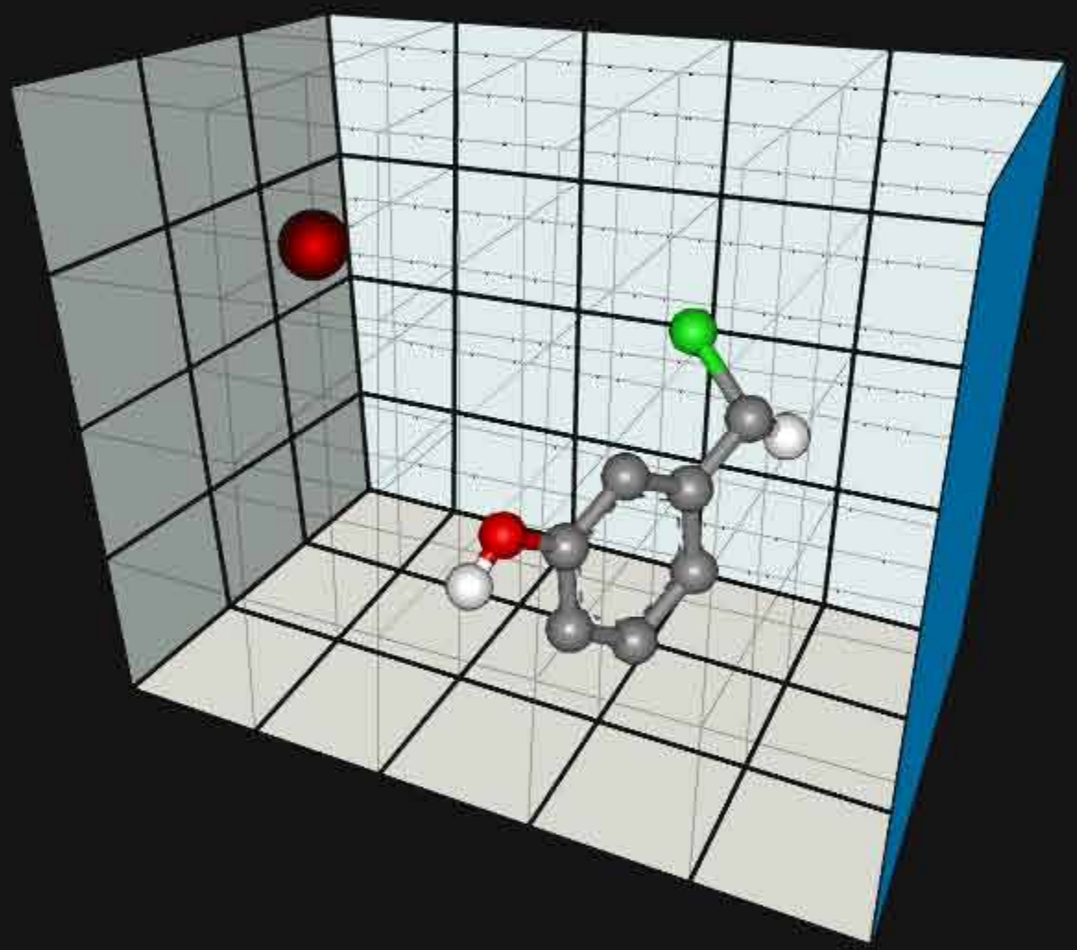
J3.5.10 3D Descriptors

Three-dimensional descriptors attempt to capture the 3D properties of the molecules to be compared, along with their relevant properties. A major difference between 3D descriptors and those of the 1D and 2D types, is that the three-dimensional properties are dependent on the particular conformation of the molecule. Thus, one can either use a low-energy structure of the molecule to calculate its three-dimensional descriptors, or run conformational sampling to explore the conformationally accessible space of the structure. While the second route introduces more information into the descriptor, it should be noted that at the same time it also introduces more noise, which does not necessarily improve the signal-to-noise ratio of the descriptor.



J3.5.11 Field-Based Descriptors

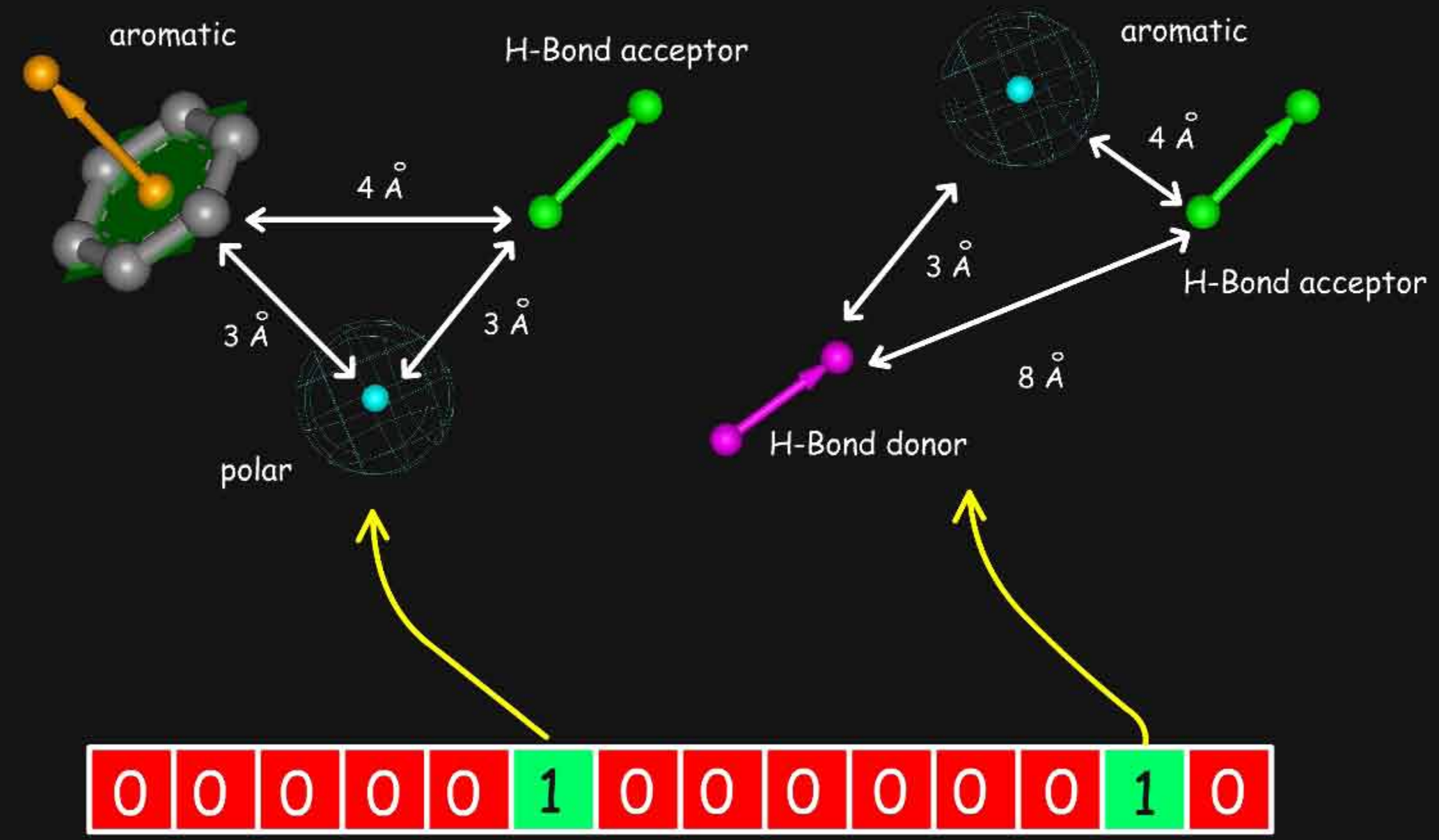
Field-based descriptors employ molecular properties calculated in space (usually on a regular grid) to represent the molecule. The best known approach is probably Comparative Molecular Field Analysis (CoMFA), which calculates steric and electrostatic properties of molecules to define both shape and surface properties which are both important for ligand-target interactions.



Steric	38.5	Kcal
Electrostatic	-9.8	Kcal

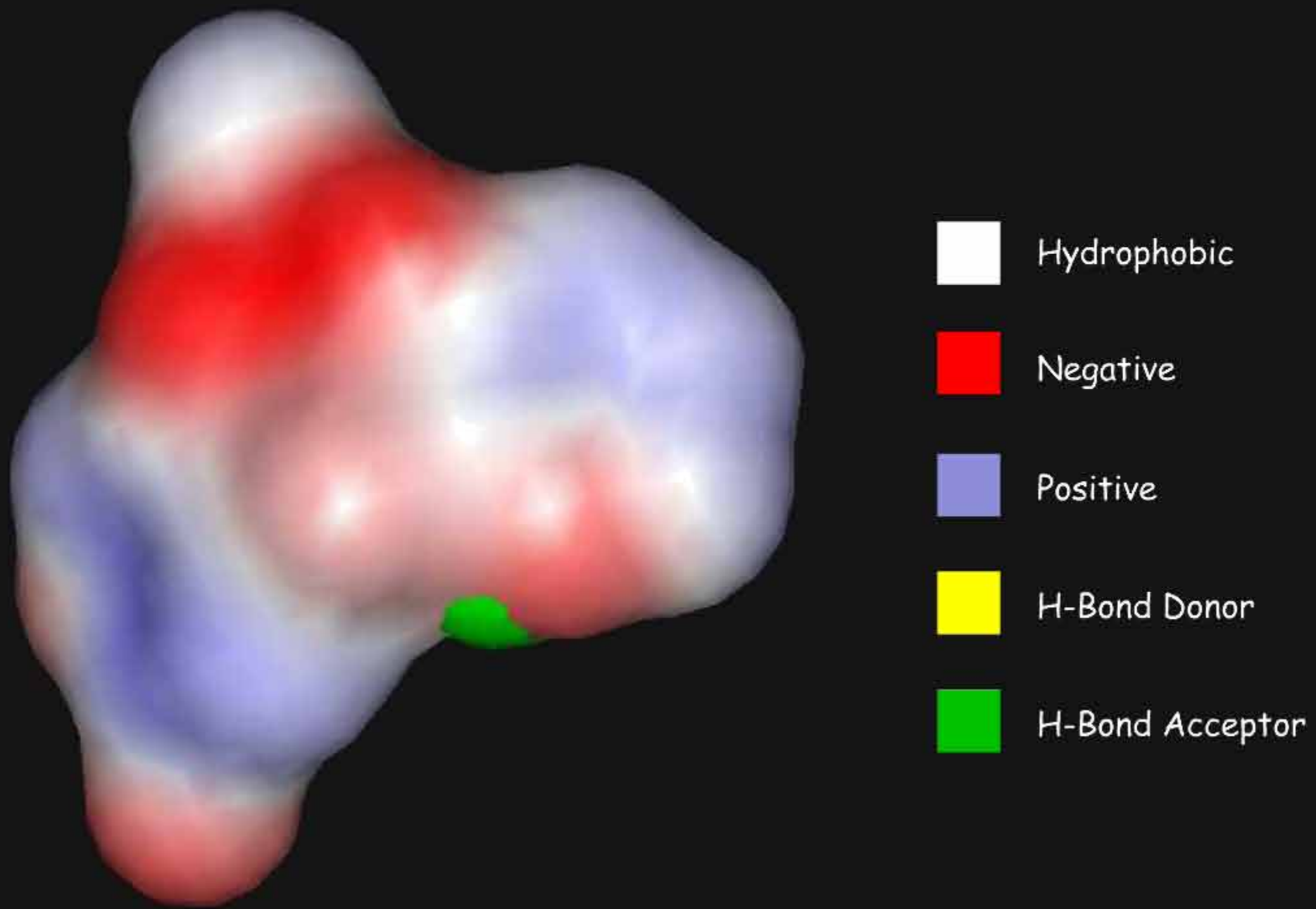
J3.5.12 Multiple-Point Pharmacophores

Multiple-point pharmacophores are generated in a two-step process: First, interaction types are assigned to individual atoms. Second, all pairs (two-point pharmacophores), triangles (three-point pharmacophores) or trapeziums (four-point pharmacophores) of the atom centers are constructed. All combinations of interaction types and distances can be assigned to bits in a fingerprint. First presented in the late 70s, this method has become very popular and represents one of the standard methods for virtual screening today.



J3.5.13 Surface-Based Descriptors

Surface-based descriptors do not employ points on a regular grid or the atom centers to describe a molecule: instead they use points on the molecular surface for this purpose. By doing this, encoding is simplified from a 3-dimensional problem in the case of grids to a 2-dimensional surface. Methods such as GRIND or MOLPRINT fall into this category.



J3.5.14 4D Chirality Descriptors

Chirality has often been neglected in fingerprint representations; however this has become an important issue in similarity searching. Fingerprint descriptors enabling the distinction between chiral molecules were first introduced by extending 3-point pharmacophore methods to 4-point fingerprints (Mason et al). More recently, other chirality descriptors have been proposed such as topological indices (Golbraith and Tropsha), and spectrum-like chirality codes (Aires-de-Sousa and Gasteiger).