Computer Aided Aroma Design. I – Molecular Knowledge Framework

Mourad KORICHI[a,b], Vincent GERBAUD[a], Pascal FLOQUET[a], A.-Hassan MENIAI[c], Saci NACEF[d], and Xavier JOULIA[a]

[a] Université de Toulouse, LGC (Laboratoire de Génie Chimique), CNRS, INP, UPS, 5 Rue Paulin Talabot, F-31106 Toulouse France

[b] LVPRS, Université de Ouargla, BP 511 Ouargla, Algeria

[c] LIPE, Université de Constantine, Constantine, Algeria

[d] LGC, Université de Sétif, Sétif, Algeria

published in Chemical Engineering and Processing

http://dx.doi.org/10.1016/j.cep.2008.02.008.

Abstract

Computer Aided Aroma Design (CAAD) is likely to become a hot issue as the REACH EC document targets many aroma compounds to require substitution. The two crucial steps in CAMD are the generation of candidate molecules and the estimation of properties, which can be difficult when complex molecular structures like odours are sought and when their odour quality are definitely subjective whereas their odour intensity are partly subjective as stated in Rossitier's review (1996). In part I, provided that classification rules like those presented in part II exist to assess the odour quality, the CAAD methodology presented proceeds with a multilevel approach matched by a versatile and novel molecular framework. It can distinguish the infinitesimal chemical structure differences, like in isomers, that are responsible for different odour quality and intensity. Besides, its chemical graph concepts are well suited for genetic algorithm sampling techniques used for an efficient screening of large molecules such as aroma. Finally, an input/output XML format based on the aggregation of CML and ThermoML enables to store the molecular classes but also any subjective or objective property values computed during the CAAD process.

Keywords: Computer Aided Aroma Design, Molecular Framework, Molecular Graph, CML, ThermoML, odour quality, subjective property.

1. Introduction

Aroma molecules are found in a wide variety of products ranging from foods, perfumes, health care products and medicines. Either combined or alone, odour and fragrance compounds are used to induce consumers to associate favourable impressions with a given product. In some rare cases (banana / isoamyle acetate, lemon / lemonal, almond / benzaldehyde), products have one predominant component which provides the characteristic odour / notes. However, in most cases, products containing odours include a mixture of fragrant compounds where a complex competition between its components sets the mixture overall odour properties.

Odour is a complex set of intensity, perception and referential-based description into a primary note and secondary note. However, truth is that olfaction phenomenon is not yet completely understood and odour measurements are often inaccurate (Amboni *et al.*, 2000).

Recently, some aroma substances have been declassified within the European Community REACH document regulating the use of chemicals in terms of environment and toxicity and forcing industry to eventually substituting existing substances by more environment friendly and less toxic ones. Such a problem is a perfect match for the application of chemical product design especially computer aided molecular design (CAMD), where we try to find a chemical product that exhibits certain desirable or specified properties (Constantinou *et al.*, 1996, Harper and Gani, 2000). With successes among which the finding of new refrigerants in replacement of proscribed CFC components or the finding of solvent in separation processes, CAMD inverse methodology has shown that two kind of information must be handled: molecular information to describe the product and thermodynamic information that concerns the product property values under given operating conditions.

For the substitution of aroma molecules using CAMD, challenges come from the size of the aroma molecules that can also display different odour between isomers and from the subjectivity of odour properties.

The paper is organized in two sections. In the first section, we present at first issues of computer aided aroma design (CAAD) by pointing specificities to aroma substitution. Secondly, an efficient molecule knowledge framework is described to be suitable for manipulation of large molecules and compliant with the use of any property estimation method, from simple structure – property models to molecular simulation within a CAAD hierarchical multi-level methodology. In a second part (Korichi *et al.*, 2008), a framework for subjective properties is proposed to classify molecules in terms of odour properties

and is applied to the definition of molecule structure - odour relationship for balsamic primary notes and five balsamic secondary notes.

2. Computer Aided Aroma Design

Computer Aided Molecular Design is a methodology of “inverse formulation” where target property values are first set and candidate molecules are sought among existing databases or constructed to satisfy the target values. These can be formulated in terms of a single objective function with proper weighting of individual properties (Constantinou *et al.*, 1996; Gani and Constantinou, 1996) or in terms of property clusters using suitable techniques which allow a high-dimensional problem to be visualized in two or three dimensions (Shelley and EL-Halwagi, 2000; Eden *et al.*, 2004; Eljack *et al.* 2005). For aroma design, the hierarchical multi-level “generate and test” CAMD methodology is well-suited (Gani *et al.*, 1991, Harper *et al.*, 1999, Harper and Gani, 2000). However, specificities of aroma molecules must be dealt with like the inherent subjectivity of assessing an odour but also their large molecular weight, their frequent combining with other odours into mixtures or their infinitesimal isomer-like differences that may lead to different odour quality. Figure 1 presents the existing CAMD methodology and tools and the additional elements that are presented in this paper to implement efficiently a computer aided aroma design.

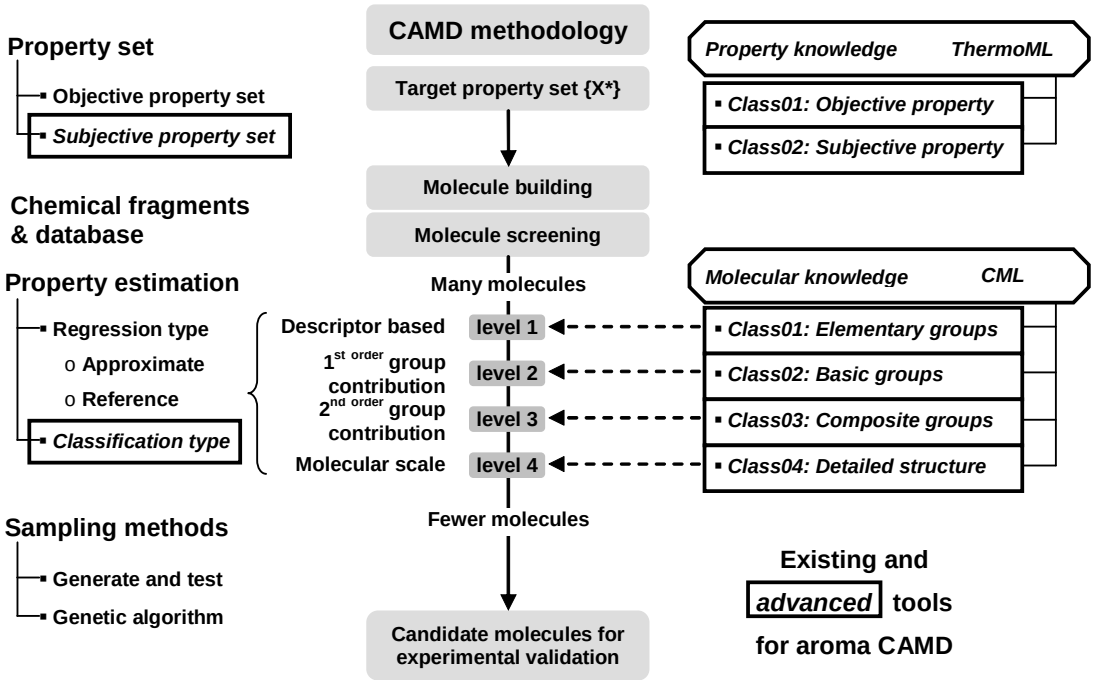


Figure 1. Computer aided aroma design framework.

The construction of a novel candidate molecule is often made from a pool of chemical fragments and proceeds hierarchically in several steps or levels at which the set of candidate is either reduced or improved based on the comparison of their estimated properties with the initial target property set.

CAMD based techniques are classified into database search, generate and test, mathematical programming and genetic algorithm (Harper *et al.*, 1999). All three can be thought of for aroma substitution but the existence of large molecules among possible fragrances hints at combining generate and test techniques with popular stochastic methods like genetic algorithm to sample efficiently the population of candidate molecules.

At each new molecule generation, physical and thermodynamic property value must be either obtained from databases or calculated from models. According to the literature, properties are classified as primary (molecular size dependent only) and secondary (dependent on the molecular structure and other variables/properties) (Gani and Constantinou, 1996). For aroma substitution, we classify in an alternate way. For objective properties, a numerical value can be measured or can be evaluated using estimation methods. Subjective properties, like odour quality or odour fondness which rely on each person's appreciation are more difficult to assess, even though referential charts have been established to harmonize the qualitative description of such properties into significant words. They are assigned a qualitative referential value. Therefore, for computer aided aroma design, we will face both real and integer type and character type properties.

Except for standard properties (normal boiling point, …), all physical and thermodynamic properties depend on the molecule surroundings conditions (temperature, pressure, solvent, …) that must also be acknowledged along with the property value. For subjective properties, rarely controllable surroundings have been known to affect the perception of the property, like painting colour, noise, wind … Such perturbations are nevertheless discarded in a CAMD methodology. Notice that for aroma, the odour intensity can be described by the odour value concept that is proportional to the ratio of the saturated vapour pressure to a threshold concentration in air. The threshold concentration has been measured for some existing fragrances but is likely unknown for novel substances (Calkin and Jellinek, 1994) and cannot be evaluated but by experiments as far as we know because it requires a panel of sensory experts even though automatic sensory devices have been tested for the black truffle aroma (Talou *et al.*, 1987; 1990). Therefore, it has a numerical value like objective properties but it is rather based on a subjective assessment of experts.

For objective type properties, according to Gani and Constantinou (1996), estimation methods are classified as "reference" (accurate but computationally expensive) and "approximate" (limited

application range but computationally simple). Models of approximate methods, established on a regression procedure, describe the chemical structure-property relationship, like QSPR methods that rely on molecular descriptor or group contribution methods that rely on the segmentation of a molecule in a set of predefined atoms or chemical groups (Reid *et al.*, 1987). Such simple methods are fast and can be used in the first levels of CAMD where candidates are numerous to discriminate them. Their drawback is the difficulty to handle key features of real molecules, like isomers, without adding complexity. At a further level in CAMD framework, when the number of candidate substances has been reduced, more sophisticated methods like multi-order group contribution methods (Marrero and Gani, 2001) or molecular simulation tools can be used to distinguish isomers, spatial conformation, or detailed reactivity compliance with some active site. Molecular simulation tools are not really estimation methods but rather belong to accurate experimental techniques ran using numerical simulations (Allen and Tildesley, 1987; Frenkel and Smit, 1996, Leach, 1996).

Using such diverse methods within the CAMD methodology is possible thanks to the hierarchical multi-level search where the number of candidate molecules is reduced at each level, thus enabling to use increasingly more sophisticated property estimation and candidate discrimination tools (Gani and Constantinou, 1996; Harper and Gani, 2000; Achenie and Sinha, 2003). Evidently, property estimation tools used at each level differ in their required input, ranging from component chemical gross formula or chemical group decomposition in group contribution methods to detailed atomic position and velocities in molecular simulation tools or spin multiplicity of ground or excited states of substances in quantum chemistry methods.

Indeed, description of candidate molecules in terms of gross chemical formula is never sufficient but for academic examples, because it is ambiguous even for simple formula (e.g. C_2H_6O can refer to ethanol or dimethyl ether). CAMD requires a deep investigation of the molecule structure. For complex problems where component mixtures are sought or isomers must be distinguished, it may lead to determine interaction between mixture components or a molecule spatial configuration. At the chemical process level, even though it is theoretically possible to gather all information on any process level by solving the equations of motion on the molecular level using molecular simulation techniques, a practical storage of the information at any level is welcome (Mangold *et al.*, 2002).

Even though authors have proposed automatic decomposition algorithm of molecules into groups, there exists no molecular framework handling all levels of description; from the gross chemical formula via various functional groups to the spatial 3D atomic coordinates; and thus suitable for input to any property estimation technique accessible at any level of the hierarchical multi-level CAMD

methodology. Discarding voluntarily mixtures, we propose in this paper a molecular class knowledge framework handling all the information needed on a single molecule at the various levels of a CAMD methodology. The molecular class knowledge framework is based on a molecular graph representation that enables reliable and efficient parsing and reorganization compatible with a CAMD framework based on genetic algorithm, even for the large molecules that we handle (Ourique and Teles, 1998). To be conveniently edited and reusable, it is stored using a CML formalism, derived from extended modelling language (XML). The physical and thermodynamic property values and their surrounding conditions are also stored into a modified ThermoML formalism, compatible with the CML formalism storing the molecular information. So the complete XML input / output formalism stores all the information needed for the CAMD problem. Variables are typed to manage both character type subjective properties and numerical type objective properties.

Group contribution methods or more sophisticated tools are fine for the prediction of objective type properties that are described by a quantitative value. However, in the substitution of aroma substances we are involved in; subjective properties predominate over any objective property. Intuition states that subjective properties still rely on structural information of the molecule, but classification rules rather than regression equations must be devised to assess the subjective property of the candidate molecule. The uncertainty of the structure – odour relationship has been reasonably handled by artificial neural network in the literature for several odour classes with compounds often bearing similar chemical structure features, once pertinent molecular descriptors of the odour class are found. Multidimensional Data Analysis is another method to obtain such a classification rule. Both are used and compared in part II to obtain the SOR of the balsamic odour with a further attempt to classify within the balsamic note the molecules into secondary odour note, as the aroma industry requires such information.

3. Multiclass molecular knowledge framework

A molecular framework handling all levels of description required within a hierarchical multi-level CAMD methodology means that information should be expanded or contracted at will from the gross chemical formula to the spatial atomic coordinates. The information should also match the prescribed input of any property estimation technique so we chose a matrix framework based on chemical graph theory and chemical knowledge even though other line based representation are available and more suitable to write. In this section, we first describe the four molecular classes suitable to describe a molecule at any of the levels of detail required by property estimation methods. Second, the automatic

decomposition of molecules into input structures is explicated. Thirdly, used for perennial storage in the compound databases, the CML-based input/output formalism of the matrix-based molecule representation and molecular knowledge classes is presented, along with the ThermoML formalism used to store the physical and thermodynamic property values and their surrounding conditions. Finally, a brief example of application is detailed.

3.1. *Molecular representation*

A molecule can be described by many useful chemical representations. From the quantum point of view, molecules are neutrons and protons surrounded by spinning electron that occupy precise energy levels corresponding to molecular orbitals. Bond segments are merely a convenient artificial representation of the existing electronic density between two atomic nuclei, related to the occupation of so-called bonding molecular orbitals concerning the nuclei. However, for CAMD, bond representation is essential and is at the core of any usable molecule representation. The molecule description must first facilitate the manipulation of its structure, second not be limited by molecular size or topological complexity, and finally, not violate elementary rules like the chemical valence rules. In general, two molecular representation families are widely used in chemistry:

- Line based extensions rely on atomic species and adjacency list. The line notation of a compound is an ordered list of symbols representing atoms, bonds and charge. There are several line notation methods available, such as the SMILES 'Simplified Molecular Input Line Specification' (Weininger, 1988; Weininger *et al.*, 1989) and the WLN 'Wiswesser Line Notation' (Smith, 1968). As a single line structure, it is readily compatible with compound lists of existing databases.
- Graph or object oriented representation where the bonds and atoms are described by detailed features. This is more suitable for the representation and manipulation of molecular structures but is less compact than line based extension. There are various types of this representation including atom tables, bond tables and many more.

In our work, we are motivated to use a molecular graph representation, with some modifications, for three main reasons. First, the use of a molecular graph assumes the bond connection between atoms in the molecule. It is not intended to be compact but can be easily expanded to provide information that may be frequently needed in analyzing the molecular structure (connection of atoms, rings, stereo-chemical and many more). Second, it enables to use classical properties estimations techniques based on group contribution methods but also on molecular descriptors (QSPR) as these seemingly coarse techniques are improving and become a reliable alternative to group contribution

estimation techniques for Computer Aided Molecular Design (Bünz *et al.*, 1998). Furthermore, graph objects can be easily generated from SMILES, WLN line representations but they are more easily handled by the sampling tools in CAMD like genetic algorithms used to improve candidate molecules in order to satisfy the set of target properties.

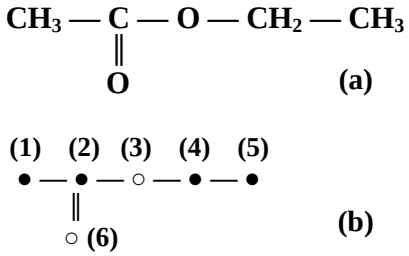
A Molecular Graph *MG = (V, E)* is defined as a mathematical representation of a molecule where atoms and bonds correspond to vertices (*V*) and edges (*E*) respectively. The MG used in this paper is undirected and unweighted and has no multiple edges between vertices (nodes) and no self-loops at any vertices, so this is called a simple graph (Pogliani, 2000). MG can be represented by a variety of matrices such as the vertex adjacency matrix, the edge adjacency matrix, the incidence matrix, the cycle matrix or the distance matrix. In this paper, the molecular structure is stored as an encoded chemical graph adjacency matrix, which is composed of the encoded atoms in a suitable type of connection table. An element of an encoded adjacency matrix cell expresses the connection relation. Zero indicates no connected elements. The proposed molecular graph (*A*) is an N x N matrix, where N is the number of atoms in the molecule backbone (Eq. 1). The off diagonal elements represent the connections between pairs of atoms. All bonds in the molecule are coded by 1. The diagonal elements (a_{ii}) represent the elementary groups present in the molecule described in section §2.2.

$$A = (a_{ij})$$

$$a_{ij} = \begin{cases} i \neq j \quad a_{ij} \begin{cases} = 1 & \text{if atom } i \text{ is connected to atom } j \\ = 0 & \text{otherwise} \end{cases} \\ i = j \quad a_{ii} \qquad \text{informations about atom } i \end{cases} \tag{1}$$

On figure 2, we present (a) the expanded chemical representation, (b) the graph and (c) the molecular graph representation for ethyl acetate. The vertices and edge sets for this molecule are:

V = {C, C, O, C, C, O} E = {(1, 2), (2, 3), (2, 6), (3, 4), (4, 5)}



$$\begin{bmatrix} a_{11} & 1 & 0 & 0 & 0 & 0 \\ 1 & a_{22} & 1 & 0 & 0 & 1 \\ 0 & 1 & a_{33} & 1 & 0 & 0 \\ 0 & 0 & 1 & a_{44} & 1 & 0 \\ 0 & 0 & 0 & 1 & a_{55} & 0 \\ 0 & 1 & 0 & 0 & 0 & a_{66} \end{bmatrix} \quad \textbf{(c)}$$

Figure 2. Ethyl acetate expanded chemical (a), graph (b) and molecular graph (c) representations.

3.2. *Automatic decomposition algorithm of molecule (ADAM)*

Automatic decomposition of molecule into groups is an important step for the prediction of physicochemical properties from diverse estimation techniques based on molecular descriptors, like Group Contribution (Joback and Reid, 1987; Constantinou and Gani, 1994, Marrero and Gani, 2001, Jaubert and Mutelet., 2004) and Quantitative Structure Property Relationship (Bünz *et al.*, 1998). In addition, it might prove useful for process simulation when molecules not yet integrated in the simulator database are incorporated in the process flows.

A few publications have concerned the automatic decomposition of molecules into functional groups used as input for property estimation packages based on the line notation, that is widely used in chemistry (Qu *et al.*, 1992a; Rowley *et al.*, 2001). Qu *et al.* (1992b) presented a decoding system suitable for a group contribution method starting from the AES (advanced encoding system) line notation technique. The AES is a modification version of WLM notation (Qu *et al.*, 1992a). Rowley *et al.* (2001) developed an automated pattern-matching program to parse the molecules into groups starting from the SMILES notation. In both research publications, two major parts are developed to ensure the decomposition of the molecules; the representation of the molecules based on line notation techniques and the parsing algorithm to search and scan the line notation. In these works, we notice primarily the absence of detailed information about 3D structure and the difficulty to cope with the advent of new property estimation techniques as the proposed decomposition are specific to a estimation method.

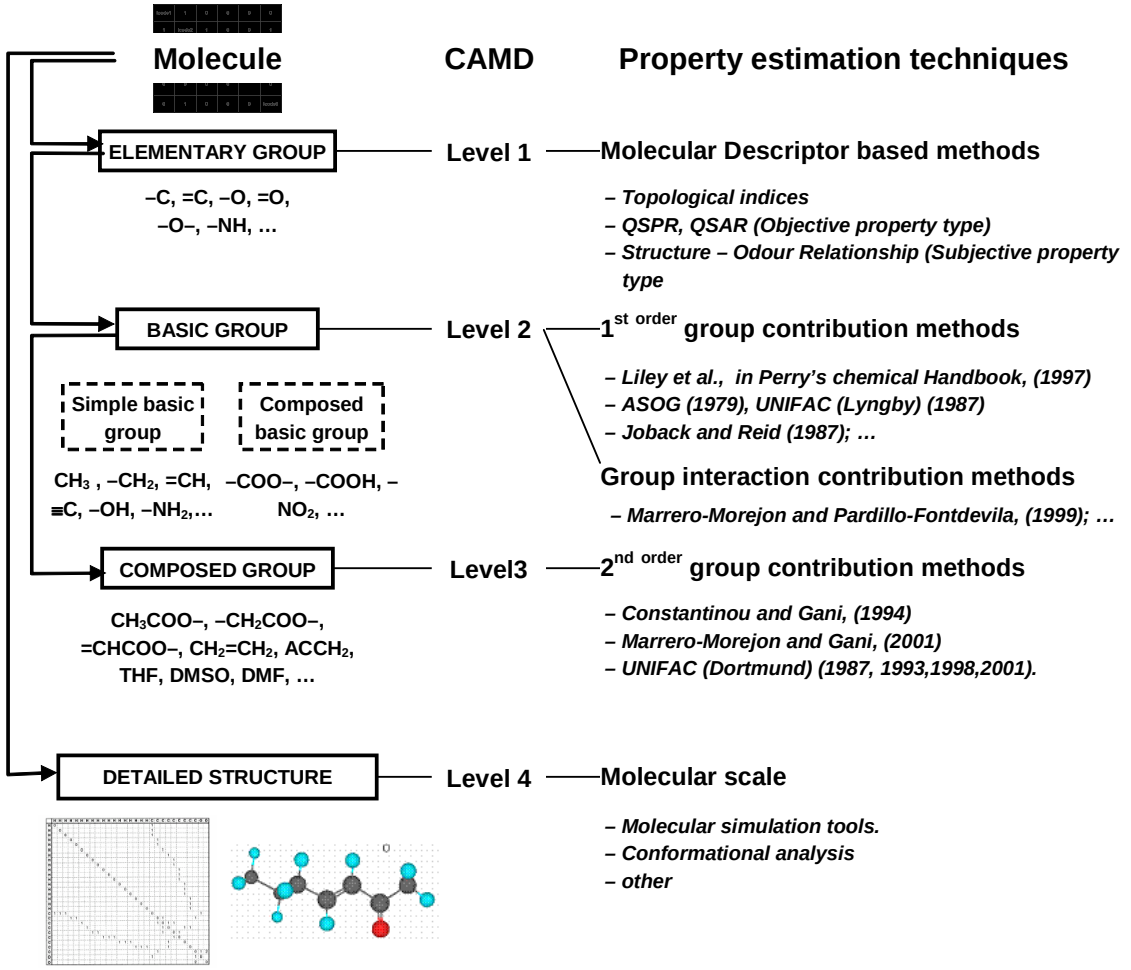


Figure 3. Molecular framework vs. estimation techniques in multilevel CAMD methodology.

Figure 3 shows how the proposed molecular knowledge framework is implemented in the hierarchical multi-level CAMD and how more and more detailed information about the molecule is used with more and more advanced property estimation methods. Figure 4 shows the three kinds of groups defined in this paper, namely elementary, basic and composed groups. Going from elementary to basic to composed groups is reversible as this proceed solely by expansion of the molecular graph structure that contains all the molecule information but its 3D structure. Decomposition into groups useful for a particular group contribution method is done via a transformation subroutine. The number of such subroutine that can be implemented within the CAMD framework is not limited and shows the evolving capacity of the approach.

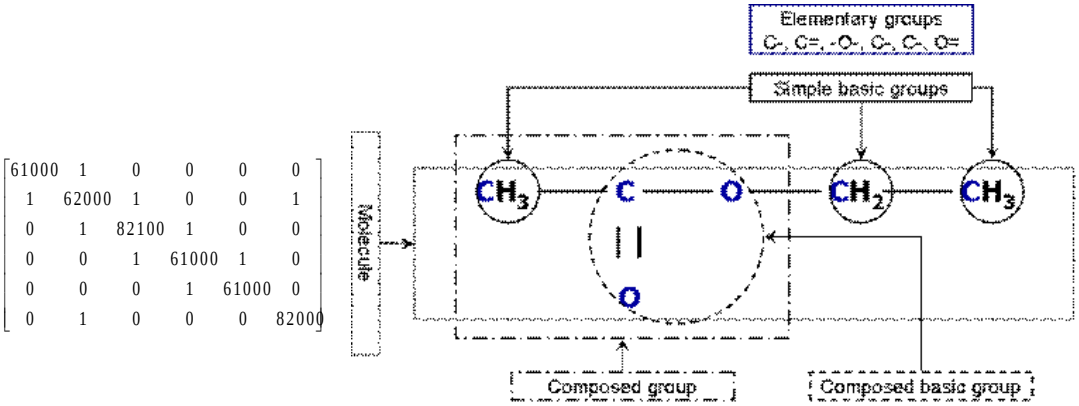


Figure 4. Molecular framework for ethyl acetate and its molecular graph.

3.2.1. Class 01: Elementary class

The elementary class represents molecules by a simple graph and identifies the connection between atoms, but with hydrogen atoms excluded. The atoms and their neighbour's information (bond, rings, position …) are described by elementary groups EG_i coded as vectors in each diagonal element (a_{ii}) of the molecular graph.

Starting from the molecular graph describing the backbone atoms and connectivity, elementary groups are generated at the first CAMD level suitable for use with molecular descriptor based methods, like topological indices, QSPR, QSAR for objective property methods or subjective property classification method like the one presented afterwards.

Each sub-element in the elementary group vector EG_i contains the following information:

$$EG_i = \{ P_1 P_2 \dots P_N \} \quad (2)$$

- Atom identifier (P_1): hydrogen atoms are excluded. For other atoms, P_1 is its atomic number. In practice, carbon, oxygen, nitrogen and halogens, sulphur and phosphorus are the most commonly encountered atoms in CAMD industrial problems and thus in existing property estimation methods.
- Bond identifier (P_2): Bonds with hydrogen atoms are not considered. Only bonds between carbon – carbon, carbon – non-carbon and non-carbon – non-carbon atoms are used. Single, double, triple and fourth bond are identified by P_2 number code 1, 2, 3 and 4 respectively. The bond identifier is independent of the connection type in the molecular structure (ring, non-ring …).
- Octal identifier (P_3): It enables to respect the octal rule and is used in equation (Eq. 2) to calculate the number of hydrogen attached to the central atom in the elementary groups.

- Ring/Non ring identifier (P_4): sets the participation or not of the atoms to an atom ring like in aromatic molecules or in glucosides. P_4 takes the value 0 if the atom is non-ring, 1 for non-aromatic rings, 2 for aromatic rings and 3 for mixed rings. For other special cases, $P_4 = 4$. This can help to specify the type of connection between atoms in molecules (carbon – non-carbon, non-carbon – non-carbon and non-carbon – non-carbon atoms).
- Others identifiers ($P_5 \ldots P_N$): The user can add any information to identify some specificity to the elementary groups. For example an OH in primary or secondary position in the structure, the number of electron in superficial layer of central atoms in the elementary group, asymmetry, an optical isomer centre, cis/trans centre, tautomery behaviour and many more.

Table 1 presents the codification of the carbon and oxygen elementary groups.

Table 1. Example of codification for carbon and oxygen elementary groups.

Atom	Chemical representation	Elementary group Codes		Observations
Carbon (C)	C-	61000	a1	Non Ring
		61010	a2	Ring
		61040	a3	Others (C-O, C-N, C-P, C-S, …)
	C=	62000	a4	Non Ring
		62010	a5	Ring
		62020	a6	Aromatic Ring
		62040	a7	Others (C=O, C=N, C=P, C=S, ...)
	C≡	63000	a8	Non Ring
		63010	a9	Ring
		63040	a10	Others (C≡N, …)
	=C=	64100	a11	Non Ring
		64110	a12	Ring
		64040	a13	Others (=C=O, =C=N, =C=P, =C=S, …)
Oxygen (O)	O=	82000	a14	Non Ring
		82010	a15	Ring
		82040	A16	Others (O=N, O=P, …)
	-O-	82100	a17	Non Ring
		82110	a18	Ring
		82110	a19	Aromatic Ring

3.2.2. Class 02: Basic group class

The second class involves the transformation of the first class molecular graph into a vector of basic groups. Basic groups are subdivided into two categories: simple basic and composed basic groups.

A simple basic group is a single elementary group with all hydrogen atoms attached to it, for example -CH_3, -CH_2, -OH. It is represented by a couple (X, Y), where X is an elementary group $EG_i=(a_{ii})$ and Y the number of hydrogen atoms NH_i connected to the central atom I and calculated by equation (Eq. 3),

$$NH_i = V_i^{STD} - \sum_{j=1,\, j\neq i}^{Natoms} a_{ij} - C_i^{(P2)} + C_i^{(P3)} + 1 \qquad (3)$$

V_i^{STD} is the standard valence of principal atom i (4 for carbon, 2 for oxygen …), a_{ij} the connection between atom neighbouring atoms j and i. The constants $C_i^{(P2)}$ and $C_i^{(P3)}$ are the bond and characteristic identifiers P_2 and P_3 respectively of atom i.

A composed basic group is built from several elementary groups with all hydrogen atoms. These groups are in limited number and deal with chemical functionality of the molecules: e.g. -COO-, -C=O, -C≡N and CH=C. The structure information of the composed basic groups is written according to equation (Eq. 4), where the Index_EG refers to the number of elementary groups and $(X, Y)_i$ refers to the i^{th} elementary groups that is defined above.

$$\left[Index\_EG, (X,Y)_1, (X,Y)_2, ..., (X,Y)_{Index\_EG}\right] \qquad (4)$$

Ethyl acetate (figure 4 & table 1) contains three simple basic groups: two -CH_3 and one -CH_2 described by (61000,3) and (61000,2) and the data structure is two [1,(61000,3)] and one [1,(61000,2)] respectively. The ethyl acetate molecule has also one composed basic group —COO— built from three elementary groups C=, O= and –O-. This composed basic group becomes [3,(62000,0), (82100,0), (82000,0)] (see figure 4 & table 2). Overall, the ethyl acetate data structure in terms of elementary groups is {4, [1, (61000,3)], [1, (61000,3)], [1, (61000,2)], [3, (62040,0),(82000,0),(82100,0)]}.

Basic groups are used mostly in 1st order group contribution methods to predict pure component properties (Lyman *et al.*, 1982, Joback and Reid, 1987, AIChE DIPPR, 1987, Liley *et al.* in Perry's chemical Handbook, 1997) or phase equilibrium properties (Fredenslund *et al.*, 1975, Larsen *et al.*, 1987, Jaubert and Mutelet, 2004, Vitu *et al.*, 2007) or in group interaction contribution methods (Marrero-Morejon and Pardillo Fontdevilla, 1999). Notice that for some group contribution methods like the original UNIFAC (Weidlicht and Gmehling, 1987), there are some ambiguous decomposition in terms of composed basic groups, especially concerning ethers and esters.

3.2.3. Class 03: Composed group class

In the third class, basic groups are combined to make composed groups defined as the connection at most of two complete basic groups, for example CH_3COO, CH_2NH_2 and CH_3CO. This step is called the relevant Functional Group Identification process (rFGI). The building approach is based on group profile principle and heuristic rules. Such composed groups are used in multi-order group contribution methods (Constantinou and Gani, 1994, Marrero-Morejon and Gani, 2001) and modified UNIFAC models (Dortmund UNIFAC from Gmehling *et al.*, 1998). The rFGI rules are specific to each property estimation methods.

As an illustration, we take the UNIFAC group CH_3COO-. This composed group can be formed by the simple basic group CH_3 and the composed basic group -COO-. Usually the connection between basic groups to generate the composed groups must be specified. The data structure of this class is written according to equation (Eq. 5), where BG_1, BG_2... BG_N are the data structure presented in equation (Eq. 4) (see table 2).

$$\left\lfloor Index\_BG, BG_1, BG_2, ..., BG_{Index\_BG} \right\rfloor \quad (5)$$

Table 2. Data structure of the tree kinds of groups

Composed group	Chemical representation	Basic group	Observations
CH3COO-	CH3	CH3	[1, (61000,3)]
	COO-	C=	[3, (62040,0),(82000,0),(82100,0)]
		O=	
		-O-	
	{2, [1, (61000,3)], [3, (62040,0),(82000,0),(82100,0)]}		

3.2.4. Class 04: Detailed structure class

In the fourth class, and for specific applications in product and process design, like odour industry where isomers have to be differentiated by using molecular simulation, the molecular structure information generated in classes 01, 02 and 03 can be refined to three-dimensional representations using atomic coordinates and other partial charges features, readily usable in molecular simulation packages. For molecular simulations based on a molecular mechanics, the atomic coordinates and the bond description provided in the molecular graph matrix is mandatory along with any specific parameters (Lennard Jones beads parameters, partial charges, multipolar moments...) (Frenkel and Smit, 1996) . For those based on quantum mechanics, the atomic coordinates only are required along with the total charge and total spin multiplicity for defining fundamental or excited states (Karplus and Porter, 1970; Leach, 1996). Molecular simulation packages already use conventional representation of molecules coordinates like 3Dmol files, PDB, z-matrix (Leach, 1996).

3.3. *Input / Output formalism*

The *eXtensible Markup Language* (XML) is a universal method to represent structured data according to normalized syntaxes and is suitable for the exchange and storage of data. Furthermore, it can be edited using any text editor. As an input/output to our framework, two standard XML formalisms are used, namely the Chemical Markup Language (CML) (Murray-Rust and Rzepa, 2001) for molecules and ThermoML for experimental properties (Frenkel *et al.*, 2003), property uncertainties (Chirico *et al.*, 2003) and predicted properties (Frenkel *et al.*, 2004). We exploit some essential definition in CML format, elements name like molecule, atom and bond and elements attributes like atomId, order and atomRef. Some modifications are proposed in the CML format to cope with the molecular knowledge framework we propose, such as the substitution of atomArray by other names like elementary group, basic group and composed group (see figure 5). The bondArray element if CML is kept.

During the CAAD methodology, the modified CML format is parsed to generate all information in the four classes and to generate the molecular graph matrix that is used during the molecule screening as genetic algorithm combined to generate and test routines enable to improve the set of candidate molecules satisfying the initial property set.

The screening of molecules requires to compare the estimated properties to the target set of properties. ThermoML is an XML-based approach for storage and exchange of experimental and critically evaluated thermophysical and thermochemical property data. Issued from a collaboration between major journals in the field like the Journal of Chemical Engineering Data and the Thermodynamics Research Center at the National Institute of Standards and Technology, it enables to store conveniently experimental data and sources (Frenkel *et al.*, 2003), but also the related uncertainties (Chirico *et al.*, 2003). Data in the form of equations and estimated data from any method are also considered (Frenkel *et al.*, 2004). Within our framework, we substitute the Compound block describing the molecule in ThermoML by the modified CML formalism explicated above. Furthermore, as suggested in by Frenkel *et al.*, 2004, we shall use the sPredictionMethodDescription [String] tag within the Prediction element of ThermoML to store the molecular descriptors used by any estimation method, e.g. the list and occurrence of the elementary, basic and composed groups. Finally, property value variables are typed to manage both character type subjective properties and numerical type objective properties.

An excerpt of XML file for ethyl acetate is presented in annexe with the CML elements concerning description of the molecule classes detailed above, and, under the sPredictionMethodDescription tag, the ethyl acetate basic groups used by the Joback method to predict the critical pressure. A subjective

property like the odour quality is also briefly described in the XML file with its classification rule based on the Linear Discriminant Analysis of 2D and 3D descriptors (see part II)..

3.4. *Illustration example*

Figure 5 illustrates briefly two major steps of the decomposition algorithm for 2-methyl butanoic acid. First, the graph decomposition and the graph analysis procedures are performed. Based on the nature of the chemical bonds described in the molecular graph, the molecule parsed in simple groups. Secondly, to set an input for some specific group contribution method, the specific group profile principle and heuristic rules are applied to locate different type of groups, especially the composed basic and composed basic groups.

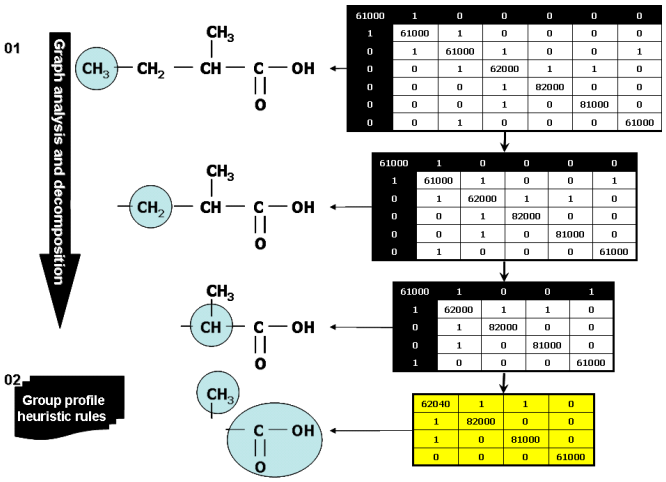


Figure 5. Decomposition procedure for 2-methyl butanoic acid.

4. Conclusions

In this paper, we present the methodology for Computer Aided Aroma Design. Specificities to aroma design are indeed the utmost importance of the subjective odour quality described with character type variables and of the partly subjective odour intensity that is the ratio of the odour threshold evaluated by a panel of sensory expert and of the fragrance vapour pressure. Besides, infinitesimal chemical structure differences like between isomers can express different odour quality and intensity. Finally aroma molecules molecular weight can be large. So, the Computer Aided Aroma Design methodology is based on the multilevel framework of the Computer Aided Molecular Design methodology but with extensions.

The multilevel of the molecular screening approach is formally matched by a molecular framework proposed in this paper. The molecular framework uses molecular graph concepts and routines to be

able at each level to provide information for evaluation using property estimation methods which complexity increases as one moves up one level. Molecular graph is particularly well suited for genetic algorithm sampling techniques that are used for an efficient generate and test sampling for the screening of large molecules such as aroma.

For each four level of the molecular screening, a molecular class is defined: Any molecule is decomposed in elementary, basic and composed groups and 3D atomic coordinates suitable for any objective property estimation methods, namely descriptor based methods, simple group contribution methods, complex group contribution methods or molecular simulation tools. Inheritance between the molecular classes enables to expand or reduce information display at will. Finally, to enhance reusability of the framework, an input/output XML format based on the aggregation of CML and ThermoML enables to store the molecular classes but also any subjective or objective property values computed during the CAAD process.

5. References

Achenie, L. E. K., and Sinha, M., The design of blanket wash solvents with environmental considerations, *Advances in Environmental Research*, 8, 213-227, 2003.

AIChE Design Institute for Physical Property Project, **Property Estimation Handbook**, American Institute of Chemical Engineers, 1987.

Allen M.P. and D.J. Tildesley, **Computer Simulation of Liquids**, Oxford University Publications, New York, Royaume-Uni, 1987.

Amboni, R. D. C., Junkes, B., Yunes R. A. and Heinzen, V. E. F., Quantitative Structure-Odour Relationships of Aliphatic Esters Using Topological Indices, *J. Agric. Food Chem.*, 48, 3517-3521, 2000.

Bünz, A. P., Braun, B. and Janowsky, R., Application of Quantitative Structure-Performance Relationship and Neural Network Models for the Prediction of Physical Properties from Molecular Structure, *Ind. Eng. Chem. Res.*, 37, 3043-3051, 1998.

Calkin RR, Jellinek JS. **Perfumery: Practice and Principles**. New York: John Wiley and Sons, Inc.; 1994

Chirico, R. D., Frenkel, M., Diky V. V., Marsh, K. N. and Wilhoit, R. C. , ThermoML - An XML-Based Approach for Storage and Exchange of Experimental and Critically Evaluated Thermophysical and Thermochemical Property Data. 2. Uncertainties. *J. Chem. Eng. Data*, 48, 1344-1359, 2003.

Constantinou L., Bagherpour K., Gani R., Klein J. A. and D. T. Wu, Computer aided product design: problem formulations, methodology and applications, *Computers & Chemical Engineering*, 20(6-7), 685-702, 1996.

Constantinou, L. and Gani, R., New group contribution method for estimating properties of pure compounds, *AIChE J*., 40, 1697-1710, 1994.

Eden M.R, Jorgensen S.B., Gani R., El-Halwagi M.M., A novel framework for simultaneous separation process and product design, *Chemical Engineering and Processing*, 43, 595–608, 2004.

Eljack F.T., Abdelhady A.F., Eden M.R., Gabriel F.B., Qin X., El-Halwagi M.M., Targeting Optimum Resource Allocation using Reverse Problem Formulations and Property Clustering Techniques, *Computers and Chemical Engineering*, 29, 2304-2317, 2005.

Fredenslund, A., Jones R.L. and Prausnitz, J.M., Group Contibution Estimation of Activity Coefficients in Nonideal Solutions, *AIChE J*., 21, 1086, 1975.

Frenkel D., Smit B., Understanding Molecular Simulation. From Algorithms to Applications, Academic Press, San Diego, 1996.

Frenkel, M., Chirico, R. D., Diky, V. V., Dong, Q., Frenkel, S., Franchois, P. R., Embry, D. L., Teague, T. L., Marsh, K. N. and Wilhoit, R. C. , ThermoML – An XML-Based Approach for Storage and Exchange of Experimental and Critically Evaluated Thermophysical and Thermochemical Property Data. 3. Critically Evaluated Data. *J. Chem. Eng. Data*, 49, 381-393, 2004.

Frenkel, M., Chirico, R. D., Diky, V. V., Dong, Q., Frenkel, S., Franchois, P. R., Embry, D. L., Teague, T. L., Marsh, K. N. and Wilhoit, R. C. , ThermoML- An XML-Based Approach for Storage and Exchange of Experimental and Critically Evaluated Thermophysical and Thermochemical Property Data. 1. Experimental Data. *J. Chem. Eng. Data*, 48, 2-13, 2003.

Gani, R., Nielsen, B. and Fredenslund, Aa., A Group Contribution Approach to Computer-Aided Molecular Design, *AIChE J.*, 37, 1318-1332, 1991.

Gani R. and L. Constantinou, Molecular structure based estimation of properties for process design, *Fluid Phase Equilibria*, 116 (1-2), 75-86, 1996.

Gmehling J., J. Lohmann, A. Jakob, J. Li, R. Joh, A Modified Unifac (Dortmund) Model. 3. Revision and Extension. Ind. Eng. Chem. Res. 37(12) (1998) 4876-4882

Harper, P. M., Gani, R., Kolar P. and Ishikawa T., Computer-aided molecular design with combined molecular modelling and group contribution, *Fluid Phase Equilibria*, 158-160, 337-347, 1999.

Harper, P. M. and Gani, R., A multi-step and multi-level approach for computer aided molecular design, *Computers and Chemical Engineering*, 24, 677-683, 2000.

Jaubert, J.N. and Mutelet, F., VLE predictions with the Peng-Robinson equation of state and temperature dependent kij calculated through a group contribution method. *Fluid Phase Equilibria*. 224(2), 285-304, 2004.

Jobak, K.G. and R.C. Reid, Estimation of pure-component properties from group contributions, *Chem. Eng. Commun.*, 57, 233-243, 1987.

Karplus M. and R.N. Porter, **Atoms and Molecules**, Benjamin, New York, 1970.

Kojima, K. and Tochigi, K., **Prediction of Vapor-Liquid Equilibria by the ASOG Method**, Elsevier/Kodansha, Tokyo, 1979.

Korichi M., V. Gerbaud, T. Talou, P. Floquet, A.H. Meniai, S. Nacef, Computer Aided Aroma Design. II – Quantitative Structure – Odour Relationship, *Chem. Eng. Proc.*, accepted, 2008

Larsen, B. L.; Rasmussen, P.; Fredenslund, A. A Modified UNIFAC Group-Contribution Model for the Prediction of Phase Equilibria and Heats of Mixing, *Ind. Eng. Chem. Res*., 26, 2274-2286, 1987.

Leach A.R., **Molecular Modelling: Principles and Applications**, Longmann, Harlow, Royaume-Uni, 1996.

Liley P.E., G.H. Thomson, D.G. Friend, T.E. Daubert, E. Buck. Prediction and correlation of physical properties. In Section 2 "Physical and chemical data" of **Perry's chemical Handbook**, 7th edition, Eds. R.H Perry, D.W. Don Green and J.O. Maloney, 2-337, 1997.

Lyman W.J., W.F. Reehl and D.H. Rosenblatt, **Handbook of Chemical Property Estimation Methods**. American Chemical Society, Washington, DC, 1990.

Mangold M., S. Motz and E. D. Gilles, A network theory for the structured modelling of chemical processes, *Chemical Engineering Science*, 57, 4099-4116, 2002.

Marrero J. and Gani R., Group-contribution based estimation of pure component properties, *Fluid Phase Equilibria*, 183-184, 183-208, 2001.

Marrero-Morejon J. and Pardillo-Fontdevilla E., Estimation of Pure Compound Properties Using Group-interaction Contributions, *AIChE J.,* 45, 615-621, 1999.

Murray-Rust P. and Rzepa, H. S., Chemical Markup, XML and the World-Wide Web. 2. Information Objects and the CMLDOM, *J. Chem. Inf. Comput. Sci.,* 41, 1113-1123, 2001.

Ourique, J.E. and Silva Telles, A., Computer Aided Molecular Design with simulated annealing and molecular graphs, *Computers and Chemical Engineering*, 22, Suppl. S615-S618, 1998.

Pogliani, L., From Molecular Connectivity Indices to Semiempirical Connectivity Terms: Recent Trends in Graph Theoretical Descriptors, *Chem. Rev*., 100, 3827-3858, 2000.

Qu, D., Su, J., Muraki,M. and Hayakawat,T. A Decoding System for a Group Contribution Method, *J. Chem. Inf. Comput. Sci*., 32, 448-452, 1992.

Qu, D., Su, J., Muraki,M. and Hayakawat,T. An Encoding System for a Group Contribution Method, *J. Chem. Inf. Comput. Sci*., 32, 443-447, 1992.

Reid R.C., J.M. Prausnitz and B.E. Poling, **The Properties of Gases and Liquids**. (4th Edition ed.), McGraw-Hill Book Co. (1987).

Rossitier, K. J., Structure-Odour Relationships, *Chem. Rev.,* 96,, 3201 – 3240, 1996.

Rowley, R, Oscarson, J. L., Rowley, R. L. and Wilding, W. V., Development of an Automated SMILES Pattern Matching Program To Facilitate the Prediction of Thermophysical Properties by Group Contribution Methods, *J. Chem. Eng. Data*, 46, 1110-1113, 2001.

Shelley M.D. and El-Halwagi M.M., Component-less design of recovery and allocation systems: a functionality-based clustering approach, *Computers and Chemical Engineering*, 24, 2081–2091, 2000.

Smith, E.G., Wiswesser Line-Formula Chemical Notation Method, McGraw-Hill, NY,1968.

Talou T., Delmas, M. and Gaset, A.. Principal constituents of Black Truffle (Tuber melanosporum) aroma. J. Agric. Food Chem., 35, 774-777, 1987.

Talou T., A. Gaset, M. Delmas, M. Kulifaj and C. Montant, Dimethyl sulphide : the secret for black truffle hunting by animals ?, Mycol. Res., 94, 277-278, 1990.

Vitu, S., Privat, R., Jaubert, J.N. and Mutelet, F., Predicting the phase equilibria of CO2 + hydrocarbon systems with the PPR78 model (PR EOS and kij calculated through a group contribution method). *Journal of supercritical fluids*, in press, doi:10.1016/j.supflu.2007.11.015 , 2007.

Weidlicht, U. and Gmehling, J., A Modified UNIFAC Model. 1. Prediction of VLE, hE and γ^{∞}, *Ind. Eng. Chem. Res*., 26 (7), 1372-1381, 1987.

Weininger, D., SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules, *J. Chem. Inf. Comput. Sci.*, 28, 31-36, 1988.

Weininger, D., Weininger, A., Weininger, J. L., SMILES. 2. Algorithm for Generation of Unique SMILES Notation, *J. Chem. Inf. Comput. Sci.*, 29, 97-101, 1989.

6. APPENDIX. Ethyl acetate XML molecular knowledge framework

```
<?xml version="1.0" encoding="UTF-8"?>
<XmlEntry>
<molecule id="Ethyl_Acetate">
    <molecule formula="C4H8O2">
    <molecule CAS="141-78-6">
    <molecule synonyms="Acetic acid, ethyl ester; CH3COOC2H5; Rcra waste number U112;....">
    <molecule 2Dmolfile="exist" file=" 141-78-6-2d.mol" source="http://webbook.nist.gov/chemistry/">
    <molecule 3Dmolfile="exist" file=" 141-78-6-3d.mol" source="http://webbook.nist.gov/chemistry/">
    ……
    <elementaryGroupArray>
        <elementaryGroup id="a1" atomType="6" bondType="1" hydrogenCount="0" ring="0" others="0"/>
        <elementaryGroup id="a7" atomType="6" bondType="2" hydrogenCount="0" ring="4" others="0"/>
        <elementaryGroup id="a14" atomType="8" bondType="2" hydrogenCount="0" ring="0" others="0"/>
        <elementaryGroup id="a16" atomType="8" bondType="2" hydrogenCount="1" ring="0" others="0"/>
    <elementaryGroupArray>
    <bondArray>
        <bond id="b1" elementaryGroupRefs="a1 a7" order="1"/>
        <bond id="b2" elementaryGroupRefs="a7 a14" order="2"/>
        <bond id="b3" elementaryGroupRefs="a7 a16" order="1"/>
        <bond id="b4" elementaryGroupRefs="a16 a1" order="1"/>
        <bond id="b5" elementaryGroupRefs="a1 a1" order="1"/>
    </bondArray>
    <basicGroupsArray>
        <bgroup id="bg1" elementaryGroupRefs="a1" hydrogenNumber="3" NumberinMolecule="2"/>
        <bgroup id="bg2" elementaryGroupRefs="a1" hydrogenNumber="2" NumberinMolecule="1"/>
        <bgroup id="bg3" elementaryGroupRefs="a7 a14 a16" hydrogenNumber="0" NumberinMolecule="1"/>
    </basicGroupsArray>
    <composedGroupsArray>
        <cgroup id="cg1" basicGroupRefs="bg1 bg3" NumberinMolecule="1"/>
        ……
    </composedGroupsArray>
</molecule>
<PureOrMixtureData>
    <property>
        <nPropNumber>1</nPropNumber>
        <Property-MethodID>
          <PropertyGroup>
```

```
            <Criticals>
                <ePropName>Critical temperature, K</ePropName>
                <Prediction>
                    <ePredictionType>Group contribution</ePredictionType>
                    <sPredictionMethodDescription>Joback method</sPredictionMethodDescription>
                    <PredictionMethodRef>……</PredictionMethodRef>
                    <sPredictionMethodDescription>
                        basicgroups="bg1 bg2 bg3" bgnumberinmethod="1 1 1"
                    </sPredictionMethodDescription>
                </Prediction>
            </Criticals>
          </PropertyGroup>
        </Property-MethodID>
    </property>
    <NumValues>
        <PropertyValue>
          <nPropNumber>1</nPropNumber>
          <nPropValue>523.78</nPropValue>
          <sPredictionMethodDescription>Joback method</sPredictionMethodDescription >
        </PropertyValue>
        <PropertyValue>
          <nPropNumber>1</nPropNumber>
          <nPropValue>524.20</nPropValue>
          <sPredictionMethodDescription>DIPPR</sPredictionMethodDescription >
        </PropertyValue>
    </NumValues>
    <property>
        <nPropNumber>2</nPropNumber>
        <Property-MethodID>
          <PropertyGroup>
            <Subjective>
                <ePropName>Odour Quality</ePropName>
                <Reference classification>Field of Odour</Reference classification>
                <Reference classificationRef>Jaubert et al. 1995</Reference classificationRef>
                <Prediction>
                    <ePredictionType>Structure Odour Relation</ePredictionType>
                    <sPredictionMethodDescription>LDA Korichi</sPredictionMethodDescription>
                    <PredictionMethodRef>Korichi et al., 2007</PredictionMethodRef>
                    <sPredictionMethodDescription>
                        … (here the LDA  correlation of 2D and 3D molecular descriptors)
                    </sPredictionMethodDescription>
            </Prediction>
            </ Subjective>
          </PropertyGroup>
        </Property-MethodID>
    </property>
    <NumValues>
        <PropertyValue>
          <nPropNumber>2</nPropNumber>
          <sPredictionMethodDescription>LDA Korichi</sPredictionMethodDescription >
          <ePropValue>fruity</ePropValue>
        </PropertyValue>
         <PredictionMethodCoefficients>2D and 3D descriptor values </PredictionMethodCoefficients>
    </NumValues>
</PureOrMixtureData>
</XmlEntry>
```