HAL
open science

Computer aided product design tool for sustainable chemical product development

Juliette Heintz, Jean-Pierre Belaud, Nishant Pandya, Moises Teles dos Santos, Vincent Gerbaud

To cite this version:

Juliette Heintz, Jean-Pierre Belaud, Nishant Pandya, Moises Teles dos Santos, Vincent Gerbaud. Computer aided product design tool for sustainable chemical product development. Computers & Chemical Engineering, 2014, vol. 71, pp. 362-376. 10.1016/j.compchemeng.2014.09.009 . hal-01093173

HAL Id: hal-01093173

https://hal.science/hal-01093173v1

Submitted on 10 Dec 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Open Archive Toulouse Archive Ouverte (OATAO)
OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible.

This is an author-deposited version published in: http://oatao.univ-toulouse.fr/
Eprints ID: 12221

Identification number: DOI : 10.1016/j.compchemeng.2014.09.009
Official URL: http://dx.doi.org/10.1016/j.compchemeng.2014.09.009

To cite this version:
Heintz, Juliette and Belaud, Jean-Pierre and Pandya, Nishant and Teles dos Santos, Moises and Gerbaud, Vincent *Computer aided product design tool for sustainable chemical product development.* (2014) Computers & Chemical Engineering, vol. 71 . pp. 362-376. ISSN 0098-1354

Any correspondence concerning this service should be sent to the repository administrator:
staff-oatao@inp-toulouse.fr

Computer aided product design tool for sustainable product development

Juliette Heintz[a,b,3], Jean-Pierre Belaud[a,b], Nishant Pandya[a,b,2], Moises Teles Dos Santos[a,b,1], Vincent Gerbaud[a,b,*]

[a] Université de Toulouse, INP, UPS, LGC (Laboratoire de Génie Chimique), 4 allée Emile Monso, F-31432 Toulouse Cedex 04, France
[b] CNRS, LGC (Laboratoire de Génie Chimique), F-31432 Toulouse Cedex 04, France

Keywords:
Computer aided product design
Genetic algorithm
Molecular graph
Bio-based molecule
Sustainable product design

A computer aided product design (CAPD) tool is proposed that finds mixtures matching target properties. Genetic algorithm crossover and mutation operators are completed with insertion or deletion operators adapted for side branches. A new substitution operator is devised for cyclic molecules. The mixture fitness is evaluated by a weighted sum of property performances. Molecules are represented by molecular graphs. They are split into molecular fragments which are built from polyatomic groups. Molecules or molecular fragments can be fixed, constrained or left free for building a new molecule. Building blocks are chemical functional groups or bio-sourced synthons. A specific coding of hydrogen-suppressed atoms is devised that can be used with various property estimation models where atom connectivity information is required. Illustration is provided through three case studies to find levulinic, glycerol and bio-based derivatives as substitute for chlorinated paraffin, methyl p-coumarate ester solvent and blanket wash solvent, respectively.

1. Introduction

The chemical industries are on the frontline of sustainable development due to the potential impact on the environment, health and safety of its product and process activities. Regulations such as the European REACH (REACH, 2006) and VOC (VOC, 2004) directives or the keen interest of consumers for eco-labelled products push the chemical industries to reconsider the products which they use and produce.

In Europe, the cost of registering chemicals to comply with REACH could exceed € 2.1 billion, based on about 30,000 substances (ECHA, 2012). Therefore there is a strong incentive to find substitute molecules and chemical products. New products need to obey environmental, health and safety constraints in addition to usual product and process requirements. Economists have argued that a doubly green chemistry perspective prevails among chemical industry engaged in green activities: one green for the reduction of their impacts on environment and one green for the use of renewable raw materials (Garnier et al., 2012). The first perspective is a direct transcript of the definition of sustainable growth in the founding Brundtland 1987 report. The second is the seventh principle of green chemistry (Anastas and Warner, 1998). As it should allow sustainable issues like toxicity or degradability to be met more easily, the use of bio-sourced molecules or synthons is a major stimulus when looking for a new product.

For finding a substitution molecule, the usual 'trial and error' approach seems inefficient unless high throughput screening is used. Instead, reverse engineering approaches, like Computer Aided Molecular Design (CAMD) are fit to handle several properties and to propose molecular structures matching the target values of these properties. In some cases, the problem of substituting a molecule may result in proposing a mixture. This further brings forth the challenge of computing mixture properties which may not always obey a linear mixing rule.

This paper presents a Computer Aided Molecular Design tool and its tailoring for finding alternative bio-sourced molecules and mixtures, with the help of model driven engineering (MDE)

* Corresponding author at: CNRS, LGC (Laboratoire de Génie Chimique), F-31432 Toulouse Cedex 04, France. Tel.: +33 5 34 323 651.
E-mail address: Vincent.Gerbaud@ensiacet.fr (V. Gerbaud).
[1] Current address: Department of Chemical Engineering, Polytechnic School of the University of São Paulo, Avenida Professor Lineu Prestes, 05088-900 São Paulo, Brazil.
[2] Current address: Shroff S. R. Rotary Institute of Chemical Technology, Block No. 402, At & PO – Vataria, Bharuch, Gujarat 393001, India.
[3] Current address: Prosim SA, 51 rue Ampère, Immeuble Stratège A, 31670 Labège, France.

http://dx.doi.org/10.1016/j.compchemeng.2014.09.009

concepts. The CAPD tool follows the general methodology of CAMD tool with several modifications. By using a genetic algorithm, this tool simultaneously optimizes the molecular structure of the components and their compositions in the mixture in order to best fit the desired properties at normal operating conditions set by the user.

After a section devoted to present background information related to CAMD, we describe the data structures and methods. They concern molecular representation, atom coding, fragment builder along with specific genetic operators to build or delete side chemical branches and to enhance changes in aromatic rings while keeping their aromaticity. Their implementation into a three software-component tool is then presented using MDE concepts. Three case studies are presented to illustrate some of the features of the tool: mixture search (case 1), search of a molecule with predefined bio-sourced synthons (case 2), two level search (case 3).

2. Background

Computer Aided Molecular Design (CAMD) aims at finding molecules that satisfy a set of property targets defined in advance (Achenie et al., 2003). CAMD relies upon four main concepts, namely, a molecular representation model, a set of property calculation models, a solving method and a performance criterion. Candidate molecules can be searched in a database or built from chemical groups. Their fitness is evaluated thanks to property estimation models by comparing the values of estimated property and the target property. Then they are discriminated according to their performance and either modified, kept as is or rejected, with the help of the solving algorithm. During the problem setting, in addition to the initial definition of the property target values, chemical blocks are pre-selected to be used in the molecular construction.

The CAMD problem solving method has often been tailored to a specific representation model. The early "generate and test" method was developed for a set of chemical groups that were also used by the group-contribution property estimation method (Gani et al., 1991; Constantinou et al., 1996). A vector of groups and their occurrences described candidates. However, a single vector may correspond to several isomer molecules and in this case a final step is required to generate the true molecules. To overcome this, some representation describing explicitly the group interconnections have been used: a genetic algorithm with adapted operators was used to generate polymers with a symbol string encoding (Venkatasubramanian et al., 1994), a binary representation of atom connectivity in molecules was used with a MILNP method (Churi and Achenie, 1996), an adjacency matrix was used with a simulated annealing (Ourique and Silva Telles, 1998), a graph representation was used with TABU search (Lin et al., 2005) and recently a graph-based representation issued from signature descriptors was used with a genetic algorithm (Herring and Eden, 2014). These explicit representations of molecule are fit for many kinds of property estimation methods once a routine for finding the groups or descriptors of the corresponding estimation method is provided.

Regarding the fitness of a candidate molecule, the differences between the predicted and target values of all properties are aggregated in a global objective function through either an arithmetic mean (Vaidyanathan and El-Halwagi, 1996) or a geometric mean (Del Castillo et al., 1996). The geometric mean penalizes severely the fitness when an individual property prediction/estimation method is far from target. In that way it is more discriminant than the arithmetic mean.

The evaluation of the performance of each molecule relies upon the calculation of property values that have been classified as product-properties, process-related properties and usage-related properties (Costa et al., 2006). Product attributes found desirable or undesirable by consumers belong to the latter class. For the CAMD problem, product requirements have to be translated into target property values, which have been done by using problem templates (Mattei et al., 2014a,b). Most product and process properties are usually described by group contribution methods (Joback and Reid, 1987; Constantinou and Gani, 1994; Martin and Young, 2001; Marrero and Gani, 2001, 2002; Nannoolal et al., 2004, 2007; Hukkerikar et al., 2012) or QSAR/TI topological index/QSPR methods (Veith and Konasewich, 1975; Karelson et al., 1996; Gani et al., 2005; Chemmangattuvalappil and Eden, 2013). Some environmental, health and safety (EHS) properties like R-phrase or CMR classification are described by similarity methods, relying upon the finding of specific molecular patterns in molecules (Gallenos, 2006).

The problem of designing a mixture is referred to as Computer Aided Product Design (CAPD) where individual molecules within the mixture and their composition must be found. Some CAMD methods have been extended to CAPD with an additional composition search (Klein et al., 1992; Gani and Fredenslund, 1993; Vaidyanathan and El-Halwagi, 1996; Duvedi and Achenie, 1997; Churi and Achenie, 1997; Sinha and Achenie, 2003). Overall, CAPD raises new issues compared to CAMD: firstly, more properties have to be matched, including more usage-related product properties or the mixture stability. Secondly, several mixture property models such as boiling point and flash point, exhibit non-linear mixing rules and need to be solved with built-in routines, which may increase the computation time. Thirdly, some usage-related properties may not be described by any suitable prediction model.

Several approaches have been taken to solve CAPD problem: some have performed a sequential search of each mixture components individually, before checking mixture properties, stability and composition (Gani, 2004; Conte et al., 2011; Papadopoulos et al., 2013; Mattei et al., 2014a,b), some others have done decomposition of the problem into a set of subproblems (Karunanithi et al., 2005), while some have solved the problem globally for a given application, for example polymer blends (Vaidyanathan and El-Halwagi, 1996). As part of a methodology for the design of formulated products, Gani and co-workers (Conte and Gani, 2011; Conte et al., 2011; Mattei et al., 2014a,b) have conceived the Virtual Product-Process Design Laboratory. They propose to run sequentially a design scenario within a computer aided stage: select a problem template and translate product needs into properties (Mattei et al., 2014a,b), choose an active ingredient of the product from the database, then design the solvents with their MIXD algorithm either from a pre-defined list or generated with a CAMD tool (Conte, 2010) and then add additives from another list and finally end up with the optimization of composition. To escape the computer-aided stage, a verification scenario is run with more accurate models, possibly involving model developments. An ultimate experimental validation ends the design activity. For overcoming the problem of consumer attributes not described by models, Solvason et al. (2009) combined an enumerating CAMD technique and MDOE (mixture design of experiments) technique. Illustrated with the formulation of a refrigerant mixture, they first solve a reverse formulation problem to find property relations that match user-defined attributes. Those relations are then used as target of a reverse problem aiming at finding the suitable mixture.

3. Methods and data structures

We have developed a CAPD tool, named as IBSS (Integrated Bio Sourced Search). It follows the general methodology of CAMD tools and is aimed at finding mixtures in which some molecule may bear bio-sourced fragments. The problem of finding a single molecule

is handled as a mixture with one element. The methods and data structures developed to cope with that tailoring are now presented.

3.1. Optimization problem

The CAPD problem is multi-objective since several properties must be matched. It is transformed into a single-objective problem, aiming at maximizing a global performance, *GloPerf*, described by an objective function *OF*, subject to *k* equality constraints and *i* inequality constraints on property targets *P*. It can be modelled as follows:

$$\begin{aligned} &OF = \max(GloPerf(MG_i, z_i, cond_j)) \\ &\text{s.t.} \quad P_k(MG_i, z_i, cond_j) = P_{k,fixed} \\ &P_{l,lowerbound} \le P_l(MG_i, z_i, cond_j) \le P_{l,upperbound} \\ &\text{s.t.} \quad \text{constraints on} \quad MG_i, z_i, cond_j \end{aligned} \tag{1}$$

The optimization variables are the molecular graph structure MG_i of the individual *i* mixture components, the mixture composition z_i and *j* conditions $cond_j$. The conditions, $cond_j$, affect the performance calculation by imposing conditions under which the properties are calculated.

The optimization variables can be constrained to allow the user to tailor the problem: the composition of any molecule, z_i and condition $cond_j$ can be fixed, bounded or free. For example, the user can impose mole fraction of an ingredient, specify a physical state of the molecule or define the range of operating conditions. Any molecule MG_i of the mixture can be fixed (ex. an active ingredient), sourced from a list of molecule (ex. a list of additives or solvents) or left free for optimization. In that latter case, one or more chemical fragments can be fixed or taken from a list of fragments to design the molecule (ex. to impose a renewable material derivative fragment in the molecule).

The global performance, *GloPerf*, is formulated as the product of a penalty function and of a weighted sum of *np* individual performance $PropPerf_p$ with weight w_p with respect to each property target.

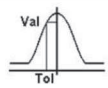
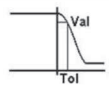
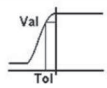
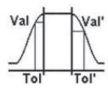
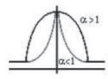
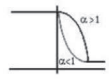
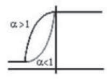
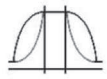
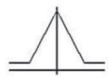
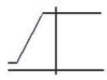
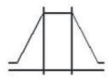
$$GloPerf(MG_i, z_i, cond_j) = \min_{\delta_r=1}(\delta_r \cdot (1 - Penal_r)) \cdot \frac{\sum_{p=1}^{np} w_p \cdot PropPerf_p(MG_i, z_i, cond_j)}{\sum_{p=1}^{np} w_p} \tag{2}$$

The penalty function *min* $(\delta_r \cdot (1 - Penal_r))$ is related to user defined rules. Each rule *r* contains data related to a molecular pattern described as an opened molecular graph and is assigned a penalty percentage $Penal_r$. δ_r is equal to 1 if the *r*th rule is violated, 0 otherwise. Typical rules describe unrealistic structures from the chemical synthesis point of view, or molecular patterns that are correlated with toxicity.

Each individual performance $PropPerf_p$ for the property *p*, compares the predicted value *x* with the targeted value *P*. The user can select among mathematical functions $F(x)$ as shown in Table 1: Gaussian (Venkatasubramanian et al., 1994), desirability functions (Del Castillo et al., 1996) or straight functions.

CAMD solution robustness is shattered by the property model prediction uncertainty. Solutions have been proposed in the literature, like the use of fuzzy logic operators to define upper and lower bounded property ranges associated to degrees of satisfaction (Ng et al., 2014), as can be done here with the straight function representation. Alternatively, the knowledge of property model uncertainty for some group contribution methods (Hukkerikar et al., 2012) can be used to define the Tol parameter in the Gaussian function representation.

Table 1
Property performance functions.

Gaussian function				
	$F(x) = G(x) = \exp\left[\ln(\mathrm{Val}) * \left(\frac{P-x}{tol}\right)^2\right]$	$F(x) = \begin{cases} 1, & x \le P \\ G(x), & x > P \end{cases}$	$F(x) = \begin{cases} G(x), & x < P \\ 1, & x \ge P \end{cases}$	$F(x) = \begin{cases} G_{inf}(x), & x < P_{min} \\ 1, & P_{inf} \le x \le P_{max} \\ G_{sup}(x), & x > P_{max} \end{cases}$
Desirability function				
	$F(x) = \begin{cases} 0, & x < P_{min} \\ D_+(x) = \left[\frac{x-P_{min}}{P-P_{min}}\right]^{\alpha}, & x \in [P_{min}; P] \\ D_-(x) = \left[\frac{P_{max}-P}{P_{max}-P}\right]^{\alpha}, & x \in [P; P_{max}] \\ 0, & x > P_{max} \end{cases}$	$F(x) = \begin{cases} 1, & x < P \\ D_-(x), & x \in [P; P_{max}] \\ 0, & x > P_{max} \end{cases}$	$F(x) = \begin{cases} 1, & x < P_{min} \\ D_+(x), & x \in [P_{min}; P] \\ 1, & x > P \end{cases}$	$F(x) = \begin{cases} 0, & x < P_{min} \\ D_+(x), & x \in [P_{min}; P_1] \\ 1, & x \in [P_1; P_2[\\ D_-(x), & x \in [P_2; P_{max}] \\ 0, & x > P_{max} \end{cases}$
Straight lines				
	$F(x) = \begin{cases} 0, & x < P_{min} \\ L_+(x) = \left[\frac{x-P_{min}}{P-P_{min}}\right], & x \in [P_{min}; P] \\ L_-(x) = \left[\frac{P_{max}-P}{P_{max}-P}\right], & x \in [P; P_{max}] \\ 0, & x > P_{max} \end{cases}$	$F(x) = \begin{cases} 1, & x < P \\ L_-(x), & x \in [P; P_{max}] \\ 0, & x > P_{max} \end{cases}$	$F(x) = \begin{cases} 1, & x < P_{min} \\ L_+(x), & x \in [P_{min}; P] \\ 1, & x > P \end{cases}$	$F(x) = \begin{cases} 0, & x < P_{min} \\ L_+(x), & x \in [P_{min}; P_1] \\ 1, & x \in [P_1; P_2[\\ L_-(x), & x \in [P_2; P_{max}] \\ 0, & x > P_{max} \end{cases}$

3.2. The search algorithm

The search algorithm selected is a genetic algorithm with elitism policy as earlier proposed by Venkatasubramanian et al. (1994) in CAMD. Modification operators are added to alter the mixture composition, conditions and molecules and to perform a multilevel search. The population size, the elitism value, the number of level and all the probabilities of operators are defined by the user.

The initial population of individuals is generated randomly within the predefined constraints on the optimization variables related to MG_i, z_i, $cond_j$. The method for building fragments from chemical building blocks is described later.

The CAPD search can be performed in several sequential levels (Harper et al., 1999; Korichi et al., 2008). At low level, simple and/or fast-computing property prediction models are used over a large population. Then as the level increments, more complex and/or time-consuming models are used over a smaller population originated from the fittest individuals of the previous level population. At the next level, the same set of building blocks and molecular structures is kept. In the meantime, the objective function can be modified according to the user's initial choices: property estimation models can be dropped, added or substituted by more complex ones.

3.3. Mixture representation data

3.3.1. Mixture representation

The mixture structure is customisable as presented in Fig. 1. Each mixture is an assembly of items and conditions. Each item contains one molecule and one mole fraction value. Each molecule is further split into interconnected fragments. The fragments are further built from basic or complex functional groups.

Initially, the user defines the mixture structure: the number of molecules, their type (fixed, list or free) and composition constraints. For each free molecule, he sets the number of fragments, fragment type (fixed, list or free) and fragment interconnections. For each free fragments, he defines the building groups to be used and their maximum number Different building block list can be used for different fragments. A molecule may contain a single free fragment. In that case the fragment has no external connections.

3.3.2. Molecular representation

We have selected molecular graph for the molecular representation which is described by an adjacency matrix (Achenie et al.,

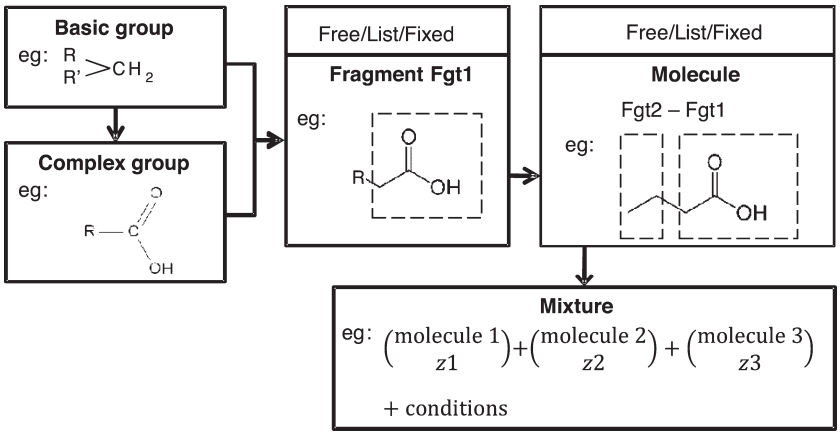


Fig. 1. Overview of the mixture structure and its substructures.

2003). The diagonal elements are either a hydrogen-suppressed atom basic group, or a complex group or a fragment. The off-diagonal elements are bond type connections. A matrix with a diagonal that contains basic groups exclusively is called an extended molecular graph hereafter.

A molecule graph is the aggregation of its fragment graphs. Fig. 2 describes the acetoin molecule (3-hydroxybutanone) built as a structure of three interconnected fragments. Fragment 1 is built from two basic groups and one complex group. Fragments 2 and 3 are just one basic group.

Basic groups (BGs) represent a hydrogen-suppressed atom, like =O, —OH, —CH_3, =CH_2, >CH_2, —NH_3, ≡N, etc. BGs are often similar to some first order groups in group contribution methods. They are assigned a "BG.ID" attribute. It is an 'elementary group' integer $EG = P_1P_2P_3P_4$ that is displayed in the extended molecular graph diagonal. P_1 refers to the atomic number preceded with a 1 (106 for C, 107 for N, 108 for O, 117 for Cl. . .), P_2 refers to the highest bond order, P_3 to the type of the atom attached to, including its occurrence in an aromatic or non-aromatic cycle, and P_4 to the number of implicit hydrogen atoms bonded to the atom. BGs also bear a "bondvector" attribute that represents the number of single, double and triple bonds. e.g. for BG.short_formula = "=C<", BG.ID = "106200" and BG.bondvector = "[2;1;0]".

Complex functional groups (CGs) are multi-atom groups. They are useful for a compact description of multi-atom chemical functions, like carboxyl groups, R—COOH, nitrite R—O—N=O, nitro R—NO_2, peroxy ROOR′, ester RCOOR′, acetals RCH(OR′)(OR″), sulfenyl RSOR′. . . A non-exhaustive list of BGs and CGs along with BG.ID is provided as Supplementary material. As they are kept intact when applying modification operators, complex groups are suitable to describe bio-sourced molecule derivative or synthons and to keep them in the molecule candidates.

CGs inherit from the basic group attributes, but the "ID" attribute has no specific meaning and is assigned a unique incremental value 2xxxxx defined by the user. Additional CGs attributes are a molecular graph "CG.graph" describing the complex group in terms of basic

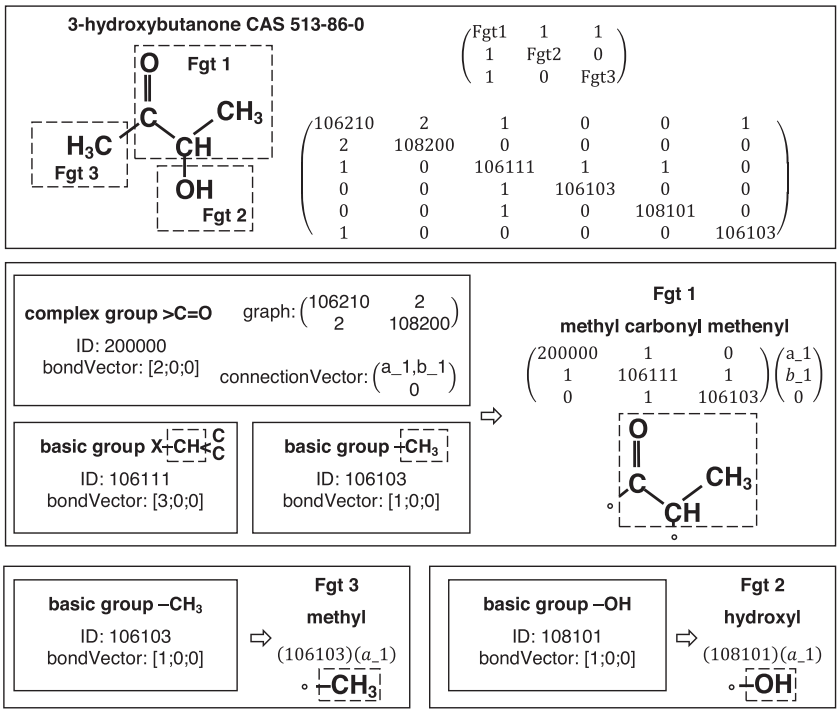


Fig. 2. Molecular graph representation of 3-hydroxybutanone as three fragments.

groups and a "CG.connectionVector" integer attribute that represents the number and location of the external connections of the group in the molecular graph (see Fig. 2).

Fragments "Fgt" are described as an adjacency matrix and an external connection vector. The adjacency matrix contains basic or complex functional group information (see Fig. 2). The vector represents the single, double or triple bond external connections, and their location on the functional groups. Similar bond type external connections on one group are distinguished by a letter.

3.4. Group vector

The classification of functional groups by Constantinou et al. (1996) is adapted into group vectors "*GV*" to provide the list of basic and complex functional groups authorized by the user to build a free fragment where k groups are allowed. A group vector *GV* is represented by the following way:

$$GV = \{N_1, N_2, \ldots, N_n\} \tag{3}$$

Where N_i is the number of groups in the fragment that have i connections, from 1 to n. Basic functional groups have up to 4 connections. Some groups with sulphur and phosphorous atoms where atom valence can be 6 and 5 respectively, though they are described with 4 connections (see a list of basic groups in Supplementary material). For complex groups, the number of connections n can be higher than 4. Authors have used some groups with 6 external connections, especially those that are sourced from renewable synthons.

In addition, the following chemical feasibility rules coming from the octet rule hold:

$$\sum_{j=1}^{n} N_j = k \tag{4}$$

$$\sum_{j=1}^{n} N_j(2-j) = 2m - \text{extconnections} \tag{5}$$

where m is a number equal to 1 minus the maximum number of cycles allowed by the user and extconnections is the number of external connections of the fragment.

3.5. Fragment creation

The method for building a free fragment from a preselected list of basic or complex functional groups is developed to ensure a diversity of structures, in particular heterocycles and aromatic cycles which may often disappear during modification by a genetic algorithm.

The procedure for constructing a fragment is described as an activity diagram in Fig. 3 which is explained below:

(1) Initialization of the user parameters: minimum $k_{\min}$ and maximum $k_{\max}$ number of BGs and/or CGs functional groups in a fragment, the maximum number of cycles m_max, the number of external connections for the fragment extconnections and the preselected BGs and CGs ordered into N_j lists.
(2) Choice of the fragment parameters: a value of k between $k_{\min}$ and $k_{\max}$ and a value of m between 1 and $1-m$_max are randomly selected. Using those values, the possible group vectors are determined, e.g. assuming that groups with up to 4 external connections exist ($n=4$) and $k=6$, five group vectors are possible $GV_1 = \{1,5,0,0\}$, $GV_2 = \{2,3,1,0\}$, $GV_3 = \{3,1,2,0\}$, $GV_4 = \{3,2,0,1\}$, $GV_5 = \{4,0,1,1\}$. A GV_i is picked and used as a basis for the fragment construction. Each group vector has the same probability to be chosen to ensure a higher diversity of the generated structures.
(3) Addition of a new element to the fragment: At this point either an acyclic group or an entire cycle structure can be added. The decision is made randomly considering the number of cycles yet to be constructed and the remaining number of elements that can be inserted. If an acyclic group is selected, a N_j is selected randomly from GV_i. Then one of the groups of the N_j *list* is picked and is added. Step 3 is repeated until all k groups have been added (go to step 6). If a cycle is selected go to step 4.
(4) Cycle building: The cycle size and its aromaticity are decided before the cycle construction starts. Then all the elements that form the cycle are inserted one after the other until the cycle is complete. Side branches to the cycle are inserted only once the cycle is closed to the cycle groups that still bear unsaturated external connections. This way the elements of a cycle are consecutive in the adjacency matrix to make the graph easier to handle.
(5) Fused cycle building: At the end of step 4 it is possible to add a fused cycle. The decision is made randomly considering the number of cycles yet to be constructed and the remaining number of elements that can be inserted. Then size and aromaticity are decided. For a non-aromatic fused cycle, the attachment points of the fused cycle are searched on the last inserted cycle and its adjacent cycles. A couple of attachment points are randomly chosen and the shortest way between these two points is determined. This path is then considered as a part of the fused cycle to build. Then the number of elements still to be added is randomly chosen and the elements are added one by one. The process to construct an aromatic cycle is the same with supplementary constraints: the attachment points must be consecutive in the last cycle; they must be connected with a double bound and must concern aromatic groups. This process goes on until the fragment is complete (go to step 6).
(6) Fragment complete: All the groups having been added, the corresponding molecular consistency is tested. If so, the possible complex groups are expanded into basic groups to create the expanded molecular graph of the fragment.

3.6. Molecule modification operators

The operators used to alter the molecular graph adjacency matrix are summarized in Fig. 4: mutation, crossover, insertion, deletion and substitution.

Regarding the insertion and deletion operators, we have added a specific branch constructor or destructor. We have also developed a new substitution operator to improve the chances of structural changes in aromatic rings without breaking their aromaticity.

- The mutation operator consists in the random replacement of a single group by a group from the N_i reference list that bears the same external connections, e.g. >NH by >CH_2 both with two single bonds in Fig. 4.
- The crossover operator consists in randomly choosing two identical non-cyclic bond types (single, double or triple bound) in two extended molecular graphs. This is checked with the P_3 value in the ID attribute of a group. The semi-graphs are then switched and recombined to form two new molecules.
- The insertion operator consists in the random addition of a group in the graph. We have developed the possibility to insert a group that has more than two connections, like —CH<, thus enabling the completion of the graph with a new branch (Fig. 4).
- The deletion operator consists in the random removal of a group with at least two connections of the same type in the graph. To be consistent with the insertion operator, it is possible to delete a group that has more than two external connections. This may

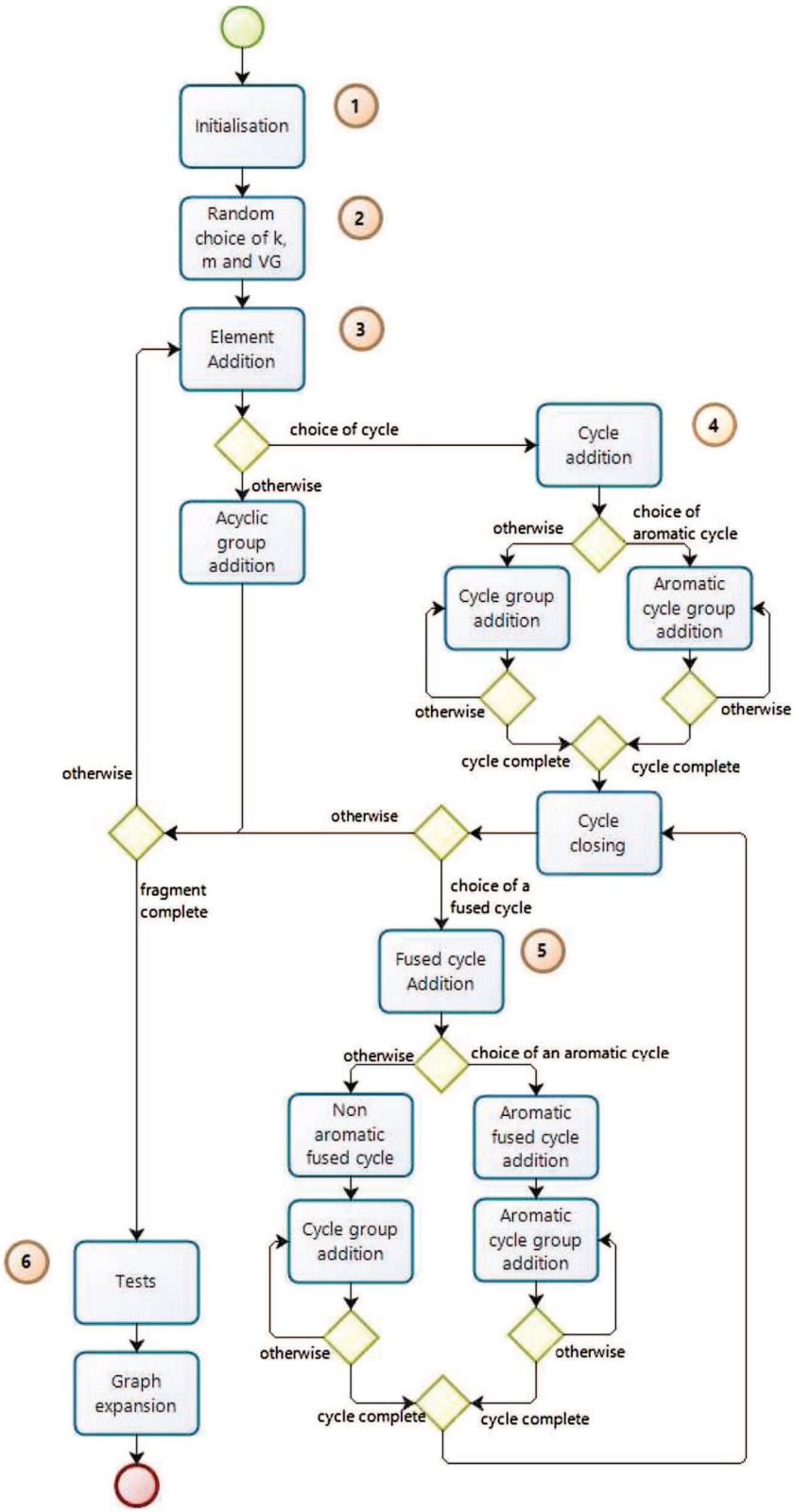


Fig. 3. Activity diagram of the creation of a free fragment object.

possibly induce the deletion of a side branch of the molecule. In Fig. 4, the group =C< is randomly chosen for deletion. The extra branches are deleted and the two remaining branches are directly reconnected. The branch NH= is deleted because it is not consistent with the remaining connection type. Thus another suitable group F— is reconnected instead.

- The new substitution operator combines the principles of mutation and insertion and consists in the replacement of a group by a group that has more connections. It was developed because the other operators performed poorly in modifying aromatic cycles without destroying their aromaticity: a mutation operator alone would be ineffective because of the limited number of aromatic

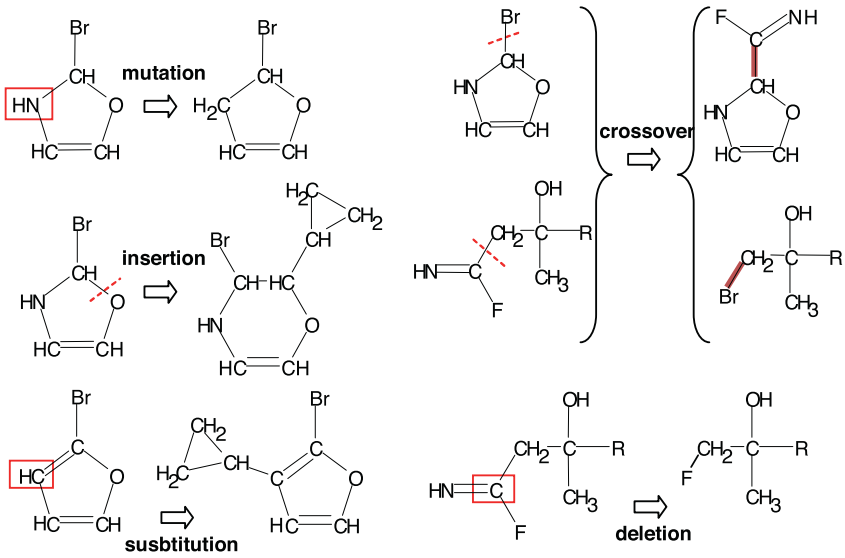


Fig. 4. Molecule modification operators.

groups, here only enabling to replace $-C_{arom}H=$ by $-N_{arom}=$ and conversely. The crossover operator cannot be applied on rings. The insertion and deletion operators would only allow adding or deleting the aromatic heteroatoms —O— and —NH—, to maintain the aromaticity.

4. Implementation

The methods described above have been implemented in the "IBSS" CAPD tool software prototype. The iterative process of the IBM-RUP software development method (Kroll and Kruchten, 2003) was applied. It is centred on the architecture and driven by the functional needs that should cover the CAPD tool. Those needs were expressed by the partners of the French ANR CDP2D 2009 project InBioSynSolv, aiming at designing biosolvents. Those needs highlighted that the sustainability of the candidate mixture may rise from the occurrence of bio-sourced fragments within the molecular structure and by the use of EHS property models to evaluate the performance of each candidate. To speed-up the implementation process, Model Driven Engineering, MDE principles were followed with the help of UML 2.0 (Unified Modelling Language) and BPMN (Business Process Modelling Notation) diagrams. It produced architectural, behavioural, functional and structural UML 2.0 views that are now briefly presented.

4.1. Architectural view

The CAPD tool is developed as component-based software. Each of the three components, 'Man–Machine-Interface MMI'; 'SEARCH' and 'P3', is a software package that encapsulates a set of functions and data and communicates through interface with the other components (Fig. 5). Next to them, an XML-structured database contains the basic and complex groups.

In the spirit of MDE, effective coding was started after the CAPD tool architecture and needs definition were well advanced. The first MMI component is written in java and aims at providing an input XML file to the search component. The second component "SEARCH" is built around an object-oriented architecture and is written in C# within the Visual Studio® environment. The third property calculation component "P3" is a Dynamic-Link Library written in VB.NET. It contains a library of property estimation models and automatic group finders or molecular descriptor routines used to translate the molecular graph information sent by the search component into suitable inputs for the property estimation models. As a standalone component, various interface methods enable the use of the P3 component with the search component or with other independent software. Currently, thirty properties can be estimated from twenty property estimation models which are listed in the Supplementary material.

4.2. Behavioural view

The behavioural view presents the different processes of the tool and highlights the components interoperability (Fig. 6).

The interoperability between the MMI and the Search components is asynchronous via an XML file as input and a text file for the results. The interoperability between the Search and the Property calculation components is synchronous and Windows-Library like. The XML file is generated through the MMI-component and loaded in the search component interface. Then, the Resolution package launches the search algorithm, a genetic algorithm with a single level in Fig. 6. First the initial population is generated. Then properties are evaluated by the property calculation component to calculate the performance of each mixture. The population is modified and evaluated again until a stop criterion is satisfied putting an end to the search. Results are then saved in a text format to the MMI package of the search component. The data in these files are afterwards displayed by the MMI component.

4.3. Functional view

The functional view highlights the potential users and the main functionalities of the software, like the function "launching a CAPD search" represented in Fig. 7. In compliance with the hierarchical decision making process for sustainable product design presented elsewhere (Heintz et al., 2014), we distinguish the expert user from the basic user. The former can access all parameters. The role of basic user is intended for people with moderate expertise in property estimation and chemistry, typically business or technical managers. The basic user has access to following functionality of the tool: to select pools of chemical building blocks classified by raw material sources, to choose product requirements to be considered, to define the rough structure of the product mixture and to set generic product constraints, like a fixed ingredient within the mixture.

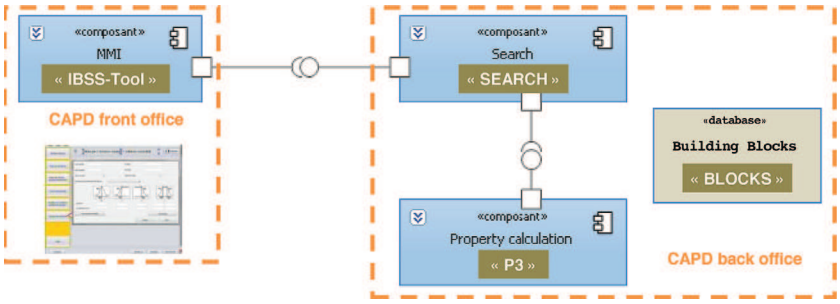


Fig. 5. CAPD tool UML component diagram.

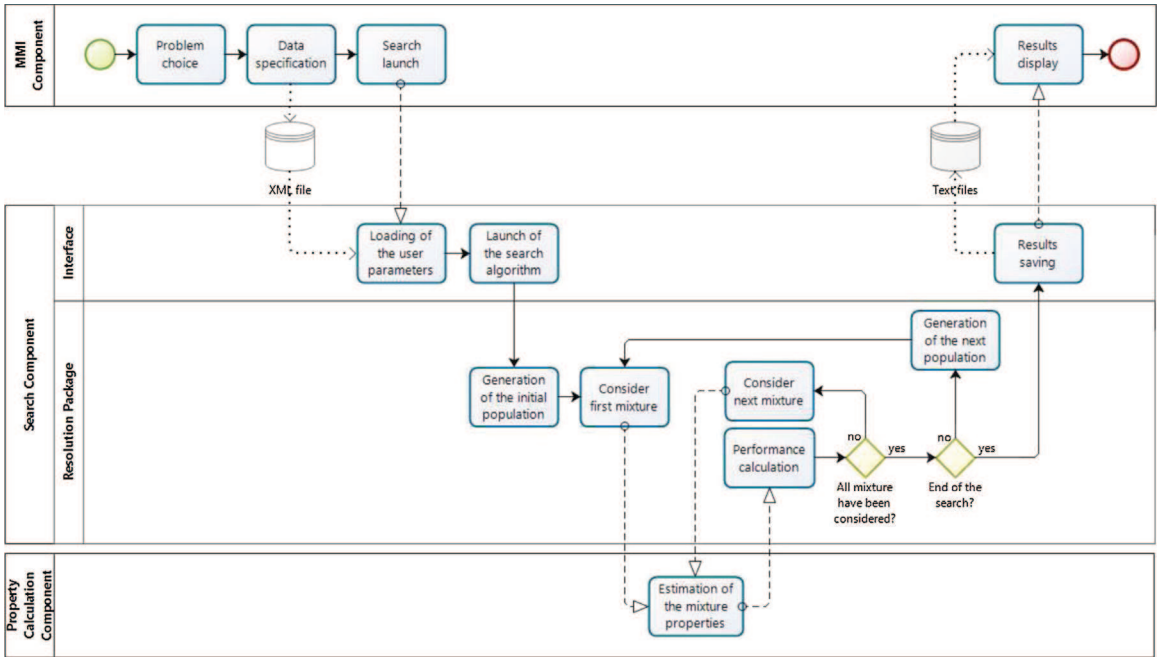


Fig. 6. BPMN diagram of the three IBSS components behaviour.

Fig. 7 describes the use case diagram of the 'launch a search' functionality. It gives access to the definition of the problem data through three data subsets and describes actions available to the user.

- The mixture data are relevant to the structure of the mixture and its components: building blocks and compositions. With these parameters, the user can customize the mixture by defining the possible fixed parts and the degrees of freedom of the different variable parts.
- The objective function data are related to the properties, their target values, their estimation models and the conditions used to calculate these properties.
- The genetic algorithm search parameters are data that can directly influence the speed and the effectiveness of the search: population size, elitism, modification operator probabilities, search level.

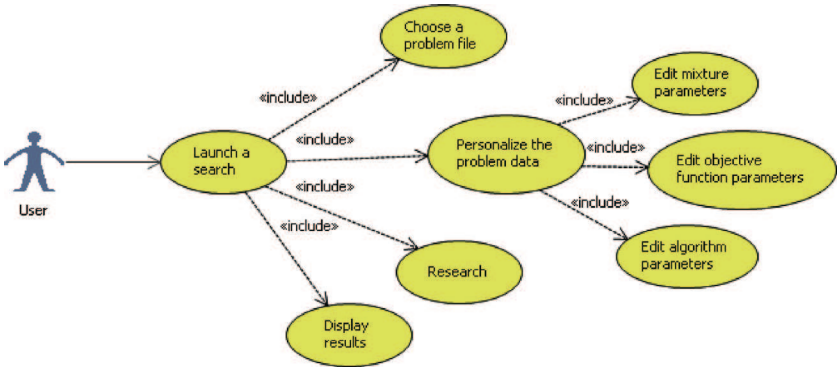


Fig. 7. "Launch a search" UML use case diagram.

Table 2
Examples of product requirements and associated calculable properties.

Product requirement	Calculable property	Default pure compound calculation model
Fluidity	Viscosity	Conte et al. (2008)
	Molecular weight	Atomic weight summation
Volatility	Boiling point	Marrero and Gani (2001)
	Vapor pressure	Riedel (1954)
	Log(Kow)	Marrero and Gani (2002)
Toxicity	−Log(LC50)	Martin and Young (2001)
	Log(BCF)	Veith and Konasewich (1975)

The mixture data should preferably be set before the objective function data because some property calculation models must be chosen for each mixture component which number must be known beforehand. The algorithm data can be specified at any time. Any data set can be saved and retrieved independently. When the data are correctly defined, the application offers the user to run the search algorithm. The results are displayed as a list of product candidate which the user can choose to save.

4.4. Structural view

The structural view concerns the model abstractions, object classes and their relationships.

- The mixture data structure implements the hierarchy described earlier in Figs. 1 and 2: mixture > composition + molecules + conditions > fragments > building blocks; along with the methods like the fragment constructor displayed in Fig. 3. A routine is developed to use the basic and complex groups stored in the "Block" database.
- The search algorithm data structure contains the genetic algorithm parameters and the modification operator methods which are described earlier.
- The objective function structure contains all the data and methods related to property target values and performance function in addition to the property estimation model for pure compounds and mixtures. In addition, we allow the basic user to be logged in the MMI component to identify product requirements as needs (Yunus et al., 2014) rather than as precise property names and models. Requirements are edited by an expert user into a set of calculable properties, with specific targets and property estimation models as done by others (Mattei et al., 2014a,b). Table 2 displays an example.

5. Case studies

5.1. Case study 1: blanket wash mixture substitution

5.1.1. Problem setting

"Blanket wash" are needed to clean ink residues from rubber blankets in the lithographic printing process. In replacement of petroleum sourced solvents, Sinha and Achenie (2003) used a CAPD approach on a pre-selection of seven water-soluble and low EHS-impacts solvents and solved a MINLP problem to find the optimal composition of the aqueous blend. Heintz et al. (2014) revisited the whole decision process leading to substitute blanket wash. It led to the specification of a tree of requirements translated here into property target values for a CAPD search with the IBSS tool. We search for an aqueous mixture and we optimize simultaneously the organic molecule and its molar composition.

A first requirement consists in solubilizing the phenolic resin ink which is evaluated by the Relative Energy Difference, RED, property given by Eq. (6). RED is computed by using Hansen solubility

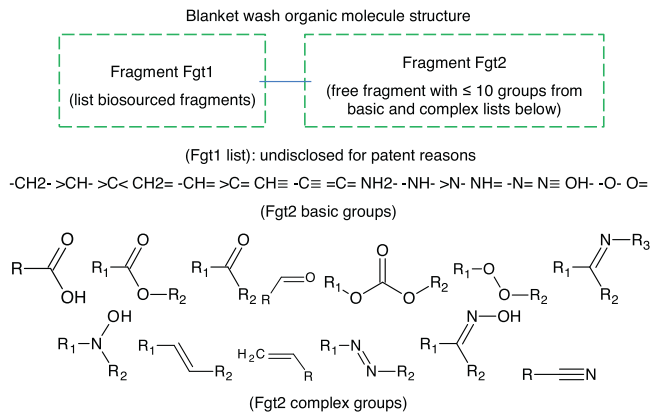


Fig. 8. Organic molecular structure and building blocks in the blanket wash mixture search.

parameters, δ_D, δ_P, δ_H, along with the Hansen distance by the ink solubility radius:

$$\mathrm{RED} = \frac{\sqrt{4(\delta_D - 19.7)^2 + (\delta_P - 11.6)^2 + (\delta_H - 11.6)^2}}{12.7} \quad (6)$$

If RED < 1, the solvent dissolves the solute. For the best solubilization results, we set a target value of RED = 0.

Other requirements refer to the cleaning process conditions: the solvent should keep its fluidity over the rubber blanket surface, which is evaluated by setting target values for viscosity and surface tension. It should comply with Volatile Organic Compound, VOC limits, evaluated with the vapour pressure. Environmental Health and Safety issues are assessed with a five EHS indices model and a flammability class is evaluated with a flash point model. The property target values are displayed in Table 3 along with the property weights, operation conditions, linear or non-linear mixture models and pure compound model. A Gaussian function is used for each property performance in Eq. (2).

In the global performance Eq. (2), we set penalties, namely the occurrence of three consecutive oxygen atoms or of C-C-C ring ($Penal_r$ = 100% => forbidden).

The organic solvent structure is split into two fragments, one with a core synthon traceable from various renewable biomass material stocks and the other built from less than 10 groups among a pool of simple and complex groups displayed in Fig. 8.

The search is ran with the following set of parameters: number of generations (300), population size (100), elitism (30) and the probabilities of crossover, mutation, insertion and deletion (respectively 65, 15, 10 and 10). It is completed in less than 40 min. Cyclic compounds are excluded. The probability for composition change and molecule change are 0.7 and 0.3 respectively.

5.1.2. Results

The output file displays a list of hundred mixtures rated by their performance (Heintz et al., 2012) and the best solution (Fig. 9) emerged after 40 generations.

Sinha and Achenie (2003) suggested a mixture of γ-butyrolactone and water (45/55 mol%) for which posterior evaluation of the performance with our criteria is 0.94. Our solution reaches a performance of 0.96 for a composition of 31 mol%. Confidential issues prevent us to display the formula of organic molecule which is different than γ-butyrolactone. Fig. 9 shows how the organic molar fraction affects the RED property. The minimum is at 0.3, indicating that the genetic algorithm correctly finds the best solution. It also shows that adding water to the biosolvent improves its cleaning ability.

Table 3
Blanket wash substitution calculable properties, target, model and parameters.

Property name	Weight	Target	Performance function[a]	Mixture model	Pure cpnd model	Operating conditions
Molecular weight	1	<200 g/mol	G(20,0.8)	–		
Flash point	1	>323.15 K	G(5,0.8)	Linear	Catoire et al. (2006)	
Vapor pressure	1	<0.00267 bar	G(10^{-4},0.8)	Linear	Riedel (1954)	T = 298.15 K
RED	4	=0	G(1,0.9)	Volumic fraction	HSPiP	See Eq. (6)
Env. waste	0.2	>8	G(1,0.8)	Linear	Weis and Visco (2010)	
Env. impact	0.2	>8	G(1,0.8)	Linear	Weis and Visco (2010)	
Health	0.2	>8	G(1,0.8)	Linear	Weis and Visco (2010)	
Safety	0.2	>8	G(1,0.8)	Linear	Weis and Visco (2010)	
LCA	0.2	>8	G(1,0.8)	Linear	Weis and Visco (2010)	
Viscosity	1	∈ [0.8;1.4] cP	G(0.1,0.8)	Non linear; Tamura and Kurata (1952)	Conte et al. (2008)	T = 298.15 K
Surface tension	1	∈ [30;45] dyn/cm^2	G(5,0.8)	Non linear; Rice and Teja (1982)	Conte et al. (2008)	T = 298.15 K
Density	1	∈ [0.9;1.1]	G(0.05,0.8)	Linear	HSPiP 3.1, Model, 2010	
Log(Ws)	4	>4 mg/L	G(0.5,0.8)	Linear	Marrero and Gani (2002)	

[a] G(Tol,Val): Gaussian type function see Table 1.

5.2. Case study 2: substitution of chlorinated paraffins

5.2.1. Problem setting

We seek an alternative molecule to chlorinated paraffins (CPs). With a general formula $C_nH_{2n+2-x}Cl_x$, they have been used as additives in high-temperature lubricants and cutting fluids for metalworking, plasticizers and flame retardants in plastics, sealants and leather (Bayen et al., 2006). Sourced from petroleum, their use and risks (toxicity, potential carcinogenic) have been assessed and are regulated (EU Report, 2008).

Table 4 summarizes the targeted physical properties.

We seek a suitable alternative sourced from levulinic acid (LA), which is known as a biomass platform chemical with derivatives already in use as lubricants, coatings and printing/inks (Bozell et al., 2000). The structure of substitute molecule consists in two fragments, one is fixed as LA and the other is constructed as assembly of less than 10 groups among 19 basic groups and 8 complex groups as displayed in Fig. 10.

Two different searches are executed: one uses only basic and complex groups to generate candidates (set 1); the other uses a fixed fragment (levulinic acid) and generate LA derivatives using basic and complex groups (set 2).

The following parameters are used: number of generations (300), population size (100), elitism (10) and the probabilities of crossover, mutation, insertion and deletion (respectively 20, 50, 15 and 15). The formation of cyclic and aromatic compounds is not allowed. No penalty related to specific chemical sub-structure is set.

5.2.2. Results

The search with set 1 (no LA fragment) produces the molecule shown in Fig. 11 that was brought out by the genetic algorithm after the 100th generation with a performance of 0.9122. It has great similarities with known CPs alternatives and displayed in Fig. 11. Notice that viscosity is not computed because of no suitable group contribution value in Joback and Reid's model exists. The user can overcome failure in property calculation by allowing the performance equation to be computed without those properties. In such situation, property weights are normalized appropriately.

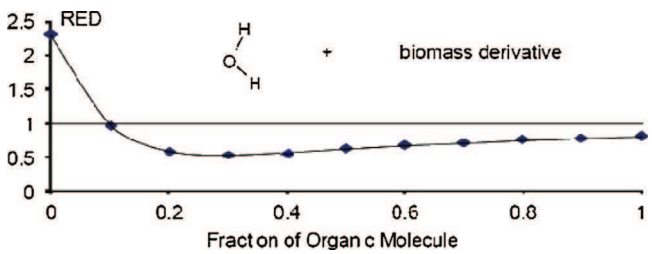


Fig. 9. Performance and property values for the best biomass derivative mixture and influence of its fraction on RED property.

Introducing LA fragment in the molecule, we display the influence of k_{max} (maximum number of building groups in fragment 2), on the performance of the genetic algorithm in Fig. 12. It shows that acceptable solutions are found if we use $k_{max} > 4$. The maximum performance (1.000) is always found for $k_{max} \geq 6$, and found more rapidly with $k_{max} = 7$.

With the set 2 (LA + basic + complex groups), the maximum performance (1.00) is achieved after 100 generations for $k_{max} = 7$. The best 10 molecules of the 100th generation are given in Fig. 13, along with a general scheme for the present approach and their predicted values of various properties in Table 5.

For the set 2, the flash point of molecules 6 and 9 could not be estimated. Indeed, group contributions do not exist for the group >CO (molecules 1–8 and 10) and for the group CHCO (molecule 9). The Configuration Interaction "CI" correction of the Hukkerikar model was not used (Hukkerikar et al., 2012). For the molecules other than 6 and 9, the viscosity was not predicted because the Joback and Reid method does not handle the phosphate group. Besides, the method of Hukkerikar et al. does not have the contribution of the group *aC-PO4* for the boiling point; therefore, the values of boiling point shown in Table 5 are only first approximations that do not take into account the *aC-PO4* group. Such property model deficiencies for complex structures confirm that CAPD computer based approaches should necessarily be validated by laboratory experiments. As mentioned before, the global performance was renormalized.

Analysing the results for molecules 6 and 9, predicted melting points values are greater than the target values but remains below the reported average absolute error of the model (17.65 K). These molecules can be retained for further experimental analysis of their melting point. Nevertheless, they will likely be waived by chemists because they exhibit a double ketone sequence, —C(=O)—C(=O)—, that is known to be unstable. Alternatively, this rule could be coded in the IBSS tool and a penalized performance could have been used instead.

5.3. Case study 3: find solvents for extraction of natural antioxidants

5.3.1. Problem setting

The objective is to find a solvent for methyl p-coumarate (MpCA), an ester of cinnamic acid isolated from plants. This class of compounds is known for their antibacterial, antifungal and/or

Table 4
CPs substitution case: product requirements and calculable properties with corresponding CAMD parameters.

Product requirement	Calculable property	Weight	Target	Performance function[a]	Pure cpnd model	Operating conditions
Liquid at usage temperature	Melting point	1	<283.15 K	G(10,0.8)	Hukkerikar et al. (2012)	
	Boiling point	1	>523.15 K	G(5,0.8)	Hukkerikar et al. (2012)	
Non flammable at high temperatures	Flash point	1	>505.15 K	G(5,0.8)	Hukkerikar et al. (2012)	
Fluidity	Viscosity	1	>37 cP	G(10,0.8)	Joback and Reid (1987)	T = 298.15 K
Low potential to bioaccumulation	Log(Kow)	1	<3	G(2,0.8)	Hukkerikar et al. (2012)	

[a] G(Tol,Val): Gaussian type function see Table 1.

Table 5
Predicted properties for the best 10 candidates for chlorinated paraffins substitution (100th generation, k_{max} = 7, levulinic + basic groups + complex groups).

	Performance	Melting point (K)	Boiling point (K)	Flash point (K)	Viscosity (cP)	Log(Kow)
Molecule 1	1.0000	274.80	358.67	518.95	–	0.31
Molecule 2	1.0000	266.34	360.24	507.30	–	0.30
Molecule 3	0.9999	283.30	362.01	520.23	–	0.36
Molecule 4	0.9986	284.72	361.60	506.98	–	−0.97
Molecule 5	0.9982	284.94	362.59	527.74	–	−0.82
Molecule 6	0.9976	284.14	317.97	–	35.19	−0.72
Molecule 7	0.9964	285.71	353.64	512.73	–	−1.45
Molecule 8	0.9902	287.39	369.86	537.69	–	0.37
Molecule 9	0.9884	280.08	307.51	–	32.38	−0.96
Molecule 10	0.9656	291.30	364.25	522.11	–	−1.33

antiviral activity (Galanakis et al., 2013; Sova, 2012). Panteli et al. (2010) found that *tert*-pentanol was the best solvent for MpCA among *tert*-butanol, *tert*-pentanol, ethyl acetate and *n*-hexane. Here we search for a glycerol based solvent. The molecular structure we look for consists in two fragments. Fragment Fgt1 is taken from a list of two glycerol derivatives with either *sn*-1 or *sn*-2 substituent. The other fragment is constructed as assembly of less than 10 groups among those displayed in Fig. 14.

We use the IBSS tool with a two level search. At level 2, we keep the three best molecules from the list of candidates found

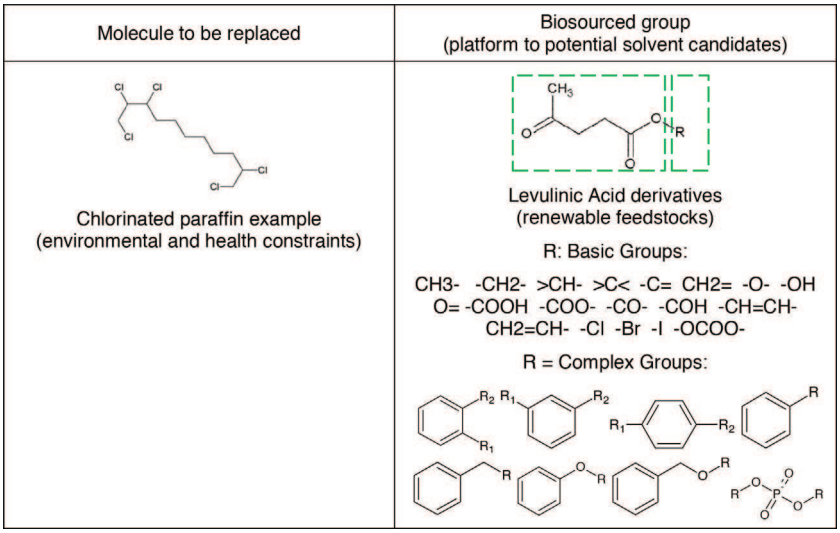


Fig. 10. A chlorinated paraffin example and building blocks used to generate alternative molecules.

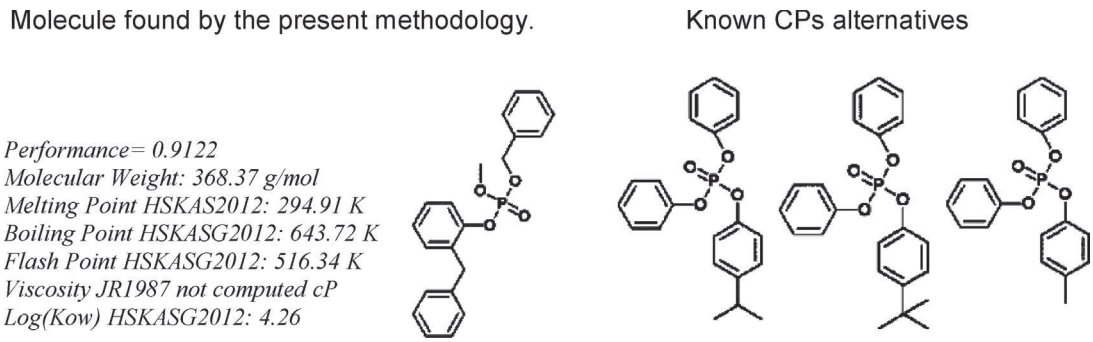


Fig. 11. Candidate and known alternatives for chlorinated paraffins (CP) substitution.

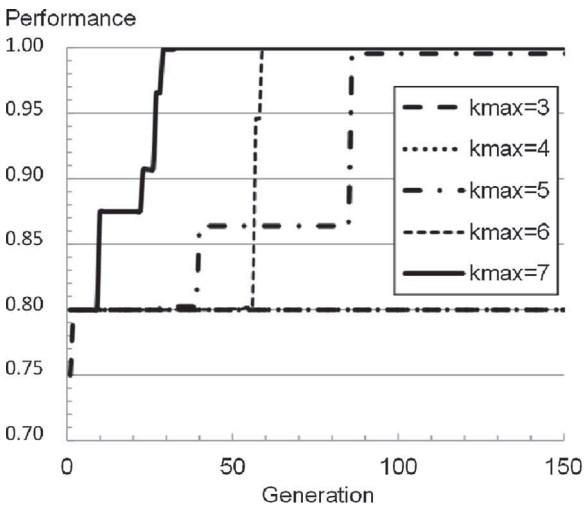


Fig. 12. Influence of the maximum number of building blocks on the genetic algorithm performance.

at level 1 and improve the prediction of their solubility by using a solid–liquid equilibrium calculation instead of the RED property (see Eq. (6)) used at level 1. The solubility x of an ideal solid phase in a liquid solvent is given by Eq. (7) in first approximation (Prausnitz et al., 1998):

$$\ln\ x = \frac{\Delta H_m}{RT}\left(1 - \frac{T}{T_m}\right) - \ln\ \gamma \tag{7}$$

where γ is the activity coefficient (computed by the UNIFAC modified Dortmund 1993 model), ΔH_m and T_m are the fusion enthalpy and the melting temperature of MpCA, respectively.

Table 6 summarizes the physical properties to be satisfied by a potential candidate.

The following set of parameters are used: number of generations (300), population size (50), elitism (5), $k_{\max}$ (4) and the probabilities of crossover, mutation, insertion and deletion (respectively 40, 50, 5, 5). The formation of cyclic and aromatic compounds is not allowed.

5.3.2. Results

The maximum performance (0.9802) is achieved in 60 generations. The best 10 molecules of the 60th generation are given in Fig. 15 and their properties are displayed in Table 7.

None of those molecules exhibit a *sn*-2 glycerol derivative. Indeed, the contribution to the melting point predictions of the *sn*-2 *OCHCH2OH* group is much larger (10.9421) than the *sn*-1 *OCH2CHOH* group contribution (1.7527) in the Hukkerikar's method (Hukkerikar et al., 2012). Thus, isomers can show a difference in melting point up to 100 K. Even though all the generated

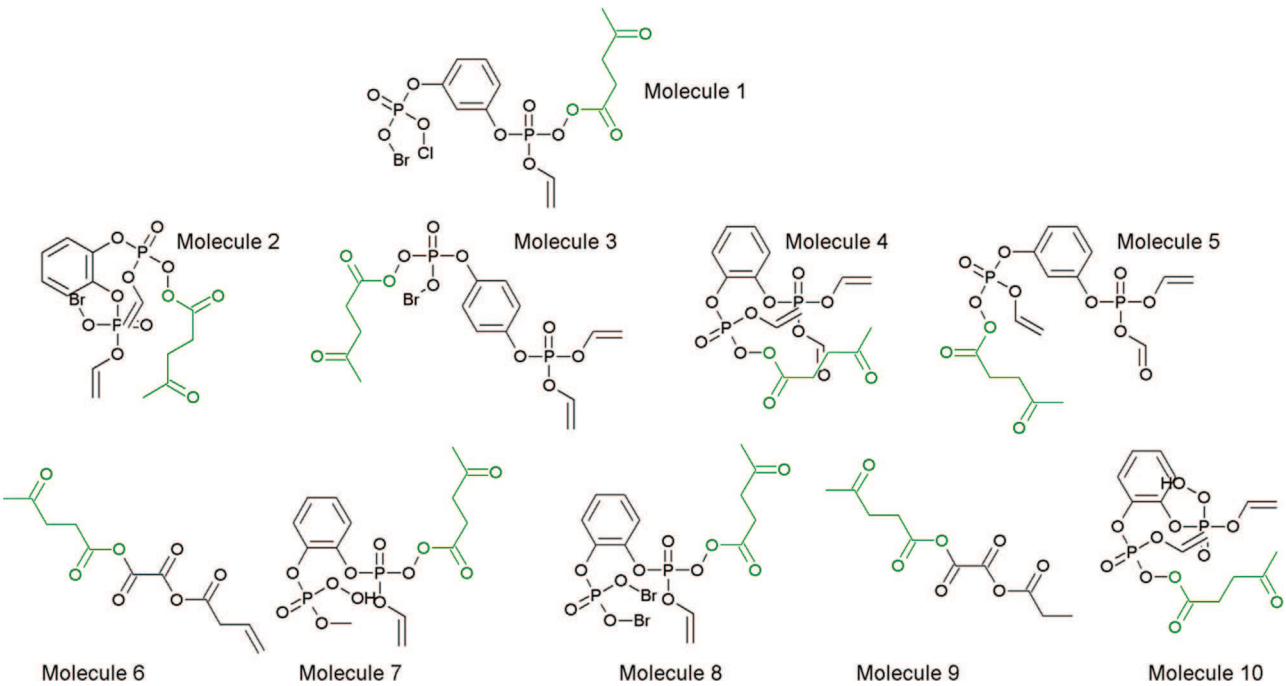


Fig. 13. Candidates for chlorinated paraffins substitution. Building blocks: levulinic acid + basic groups + complex groups. 100th generation, $k_{\max} = 7$.

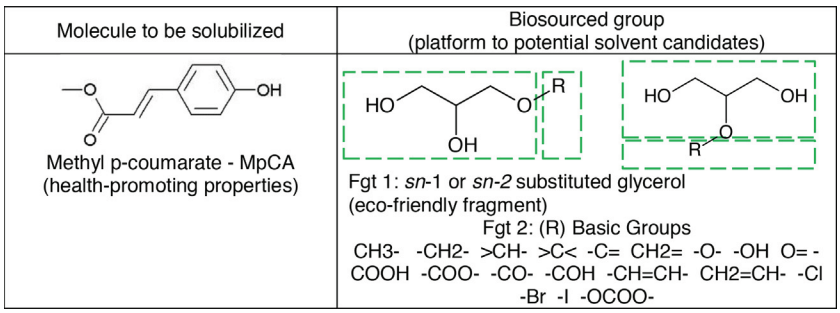


Fig. 14. Methyl p-coumarate solubilization: target molecule and building blocks.

Table 6
MpCA solvent extraction case: product requirements and calculable properties with corresponding CAMD parameters.

Product requirement	Calculable property	Weight	Target	Performance function[a]	Pure cpnd model Level 1	Pure cpnd model Level 2
Must dissolve MpCA	RED	2	=0	G(1,0.95)	Hukkerikar et al. (2012)[b]	UNIFAC + SLE at 25 °C
Liquid at room temperature	Melting point	1	<283.15	G(10,0.8)	Hukkerikar et al. (2012)	–
	Boiling point	1	>373.15 K	G(5,0.8)	Hukkerikar et al. (2012)	–
Non flammability	Flash point	1	>343.15 K	G(1050.8)	Hukkerikar et al. (2012)	–
Low potential to bioaccumulation	Log(Kow)	1	<3	G(2,0.8)	Hukkerikar et al. (2012)	–

[a] G(Tol,Val): Gaussian type function see Table 1.
[b] The Hukkerikar model is used to compute the Hansen solubility parameters, which are used to compute RED.

Table 7
Predicted properties for the best 10 candidates for methyl p-coumarate solubilization (60th generation, k_{max} = 4, glycerol + basic groups).

	Level 1						Level 2
	Performance	Melting point (K)	Boiling point (K)	Flash point (K)	Log(Kow)	RED	Solubility fraction
Molecule 1	0.9802	285.24	532.92	349.35	0.03	1.04	0.0896
Molecule 2	0.9732	287.34	548.81	341.47	1.29	0.99	0.0787
Molecule 3	0.9710	290.63	529.80	349.19	0.17	0.75	0.0902
Molecule 4	0.9635	283.62	537.89	339.35	1.07	0.99	–
Molecule 5	0.9570	293.23	523.40	345.25	0.09	0.74	–
Molecule 6	0.9471	290.25	535.61	340.53	−0.84	1.24	–
Molecule 7	0.9213	298.71	539.53	356.66	0.55	0.74	–
Molecule 8	0.9196	295.18	535.27	343.21	−1.09	1.46	–
Molecule 9	0.9150	295.13	534.63	338.28	0.78	0.67	–
Molecule 10	0.9095	300.37	556.14	347.41	0.96	0.77	–

molecules have predicted melting points greater than the target, this difference is within the reported average absolute error of the model (17.65 K), and these molecules can be retained for further experimental analysis. It can be also noted that the best molecule is a compromise, with a larger RED than other molecules but with an overall better performance.

The top 3 molecules proposed by level 1 are retained for further analysis of their solubility with the UNIFAC–SLE approach. The results of level 2 shift the order of molecules 1 and 2 in terms of performance compared to that of the level one by using the RED model: the molar fraction solubility of molecule 2 is predicted at 0.0787, lower than that of molecule 1 at 0.0896. The predicted solubility of molecule 3 is the highest: 0.0902. This value confirms with the level 1 RED prediction, in which molecule 3 is that with the lowest RED value.

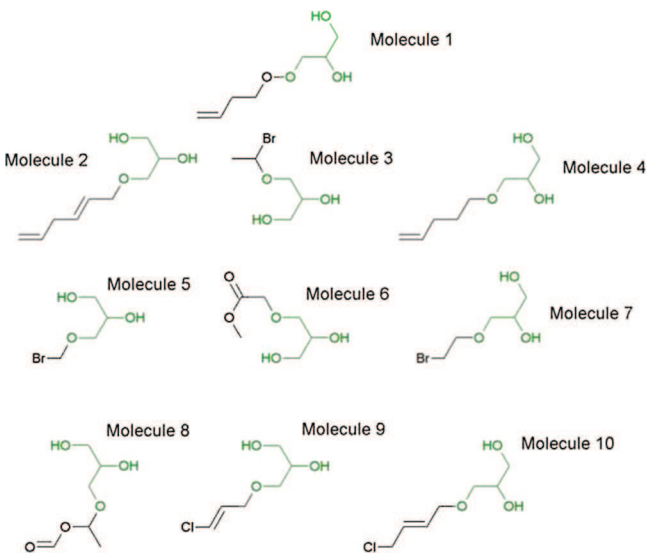


Fig. 15. Candidates for methyl p-coumarate solubilization. Glycerol and basic groups as building blocks; 60th generation.

6. Conclusion

We have described the methods, data structures and software implementation of IBSS, a computer aided product design (CAPD) tool that is aimed at finding mixtures. Based on CAMD concepts, it is able to optimize simultaneously the mixture components, compositions and mixture conditions.

With the help of a model driven engineering approach, the architecture and the functional needs of the CAPD tool were devised, and specifically aimed at finding mixtures with molecules that can be sourced from pools of renewable materials. This is done by setting constraints on the mixture molecules or on fragments within molecules, like bio-sourced synthons.

A molecular representation based on an adjacency matrix was selected. It was made flexible to represent both atom-based structures and fragment-based structures. Diagonal hydrogen-suppressed atomic elements of the matrix were described with a novel coding of four indexes describing the atom number, the highest bond type, the atom neighbouring context and the number of hydrogen atoms. The coding was also devised for the identification of chemical groups or descriptors necessary to evaluate properties with models from various authors. A complex group describing polyatomic structures was created to represent polyatomic chemical functions and synthons. It enabled to keep intact bio-sourced synthons during the search and promote their occurrence in the final solution if their performance was deemed high enough.

A genetic algorithm was selected to optimize simultaneously the mixture elements, its composition and additional operating conditions. It was made capable to perform a multilevel search with different objective functions. Modification operators were adapted to cope with the mixture search context and with the aforementioned molecular representation. Classical crossover and mutation operators were used, while insertion and deletion operators were adapted to insert or delete side branch. A new substitution operator was proposed to maintain cyclic molecules through the genetic algorithm generations. The building of fragments from basic or complex groups was made more efficient by using a vector group classification inventorying building groups on the basis of their external connections.

The mixture performance is calculated through a sum of weighted property performance that can be penalized if specific molecular patterns occur, like those found in toxic molecules.

The tool was implemented as a set of three independent software components, namely a MMI component, a CAPD search component and a property library component; associated to a database of basic and complex functional groups. To cope with the difficulty for some users of expressing requirements in terms of numerical target values, we distinguished product requirement (better suitable to basic user) vs calculable properties (for expert users).

Future work is ongoing to set the CAPD tool within a virtual laboratory decision-making framework as described by Heintz et al. (2014). We also intend to benefit from the flexible tool architecture, first by adding alternative solving methods other than genetic algorithm, and by testing the suitability of the molecular graph representation for novel property models based on higher dimensionality which are useful to handle conformation dependent properties.

Acknowledgment

This scientific work was supported by the French National Research Agency (InBioSynSolv ANR-CP2D-2009-08) in partnership with Rhodia-Solvay Company, ENSCL and LCA-INPT.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.compchemeng.2014.09.009.

References

Achenie LEK, Gani R, Venkatasubramanian V. Computer aided molecular design: theory and practice. Amsterdam: Elsevier; 2003.
Anastas P, Warner J. Green chemistry theory and practice. Oxford: Oxford University Press; 1998. p. 135.
Bayen S, Obbard JP, Thomas GO. Chlorinated paraffins: a review of analysis and environmental occurrence. Environ Int 2006;32:915–29.
Bozell JJ, Moens L, Elliott DC, Wang Y, Neuenscwander GG, Fitzpatrick SW, et al. Production of levulinic acid and use as a platform chemical for derived products. Resour Conserv Recycl 2000;28:227–39.
Catoire L, Paulmier S, Naudet V. Experimental determination and estimation of closed cup flash points of mixtures of flammable solvents. Process Saf Prog 2006;25:33–9.
Chemmangattuvalappil NG, Eden MR. A novel methodology for property-based molecular design using multiple topological indices. Ind Eng Chem Res 2013;52:7090–103.
Churi N, Achenie LEK. Novel mathematical programming model for computer aided molecular design. Ind Eng Chem Res 1996;35:3788–94.
Churi N, Achenie LEK. The optimal design of refrigerant mixtures for a two-evaporator refrigeration system. Comput Chem Eng 1997;21S:S349–54.
Constantinou L, Gani R. New group contribution method for estimating properties of pure compounds. AIChE J 1994;40:1697–710.
Constantinou L, Bagherpour K, Gani R, Klein JA, Wu DT. Computer aided product design: problem formulations, methodology and applications. Comput Chem Eng 1996;20:685–702.
Conte E. Innovation in Integrated Chemical product-process Design: Development through a Model-based Systems Approach. Technical University of Denmark (DTU); 2010, PhD Thesis.
Conte E, Gani R, Ng KM. Design of formulated products: a systematic methodology. AIChE J 2011;57:2431–49.
Conte E, Gani R. Chemicals-based formulation design. Comput Aided Chem Eng 2011;29:1588–92.
Conte E, Martinho A, Matos HA, Gani R. Combined group-contribution atom connectivity index-based methods for estimation of surface tension and viscosity. Ind Eng Chem Res 2008;47:7940–54.
Costa R, Moggridge GD, Saraiva PM. Chemical product engineering: an emerging paradigm within chemical engineering. AIChE J 2006;52:1976–86.
Del Castillo E, Montgomery DC, McCarville DR. Modified desirability functions for multiple response optimization. J Qual Technol 1996;28:337–45.
Duvedi AP, Achenie LEK. On the design of environmentally benign refrigerant mixtures: a mathematical programming approach. Comput Chem Eng 1997;21:915–23.
ECHA. REACH market final report; 2012, Available from: http://ec.europa.eu/enterprise/sectors/chemicals/files/reach/review2012/market-final-report_en.pdf [last accessed February 2014].
EU report. European Union Risk Assessment report, alkanes, C10-13, chloro. European Chemicals Bureau; 2008, http://echa.europa.eu/documents/10162/6434698/orats_addendum_alkanes_c10-13_chloro_en.pdf [last accessed September 2013].
Galanakis CM, Goulas V, Tsakona S, Manganaris GA, Gekas V. A knowledge base for the recovery of natural phenols with different solvents. Int J Food Prop 2013;16:382–96.
Gallenos SA. Mini-review on chemical similarity and prediction of toxicity. Curr Comput Aided Drug Des 2006;2:105–22.
Gani R, Fredenslund A. Computer aided molecular and mixture design with specified property constraints. Fluid Phase Equilib 1993;82:39–46.
Gani R, Nielsen B, Fredenslund A. A group contribution approach to computer-aided molecular design. AIChE J 1991;37:1318–32.
Gani R. Chemical product design: challenges and opportunities. Comput Chem Eng 2004;28:2441–57.
Gani R, Harper PM, Hostrup M. Automatic creation of missing groups through connectivity index for pure-component property prediction. Ind Eng Chem Res 2005;44:7262–9.
Harper PM, Gani R, Kolar P, Ishikawa T. Computer-aided molecular design with combined molecular modeling and group contribution. Fluid Phase Equilib 1999;158–160:337–47.
Heintz J, Touche I, Teles Dos Santos M, Gerbaud V. An integrated framework for product formulation by computer aided mixture design. Comput Aided Chem Eng 2012;30:702–6.
Heintz J, Belaud JP, Gerbaud V. Chemical enterprise model and decision-making framework for sustainable chemical product design. Comput Ind 2014;65:505–20.
Herring RH III, Eden MR. De novo molecular design using a graph-based genetic algorithm approach. Comput Aided Chem Eng 2014;33:7–12.
HSPiP 3.1. http://hansen-solubility.com/index.php; 2010.
Hukkerikar AS, Sarup B, Ten Kate A, Abildskov J, Sin G, Gani R, et al. Group-contribution+ (GC+) based estimation of properties of pure components: improved property estimation and uncertainty analysis. Fluid Phase Equilib 2012;321:25–43.
Joback R, Reid RC. Estimation of pure-component properties from group-contributions. Chem Eng Commun 1987;57:233–43.
Karelson M, Lobanov VS, Katrinsky AR. Quantum-chemical descriptors in QSAR/QSPR studies. Chem Rev 1996;96:1027–43.
Karunanithi AT, Achenie LEK, Gani R. A new decomposition-based computer-aided molecular/mixture design methodology for the design of optimal solvents and solvent mixtures. Ind Eng Chem Res 2005;44:4785–97.
Klein JA, Wu DT, Gani R. Computer aided mixture design with specified property constraints. Comput Chem Eng 1992;16S:S229–36.
Korichi M, Gerbaud V, Floquet P, Meniai AH, Nacef S, Joulia X. Computer aided aroma design I – molecular knowledge framework. Chem Eng Proc: Process Intensif 2008;47:1902–11.
Kroll P, Kruchten P. The rational unified process made easy: a practitioner's guide to the RUP. 1st ed. Boston: Addison Wesley; 2003.
Lin B, Chavali S, Camarda K, Miller DC. Computer-aided molecular design using Tabu search. Comput Chem Eng 2005;29:337–47.
Marrero J, Gani R. Group-contribution based estimation of pure component properties. Fluid Phase Equilib 2001;183–184:183–208.
Marrero J, Gani R. Group-contribution-based estimation of octanol/water partition coefficient and aqueous solubility. Ind Eng Chem Res 2002;41:6623–33.
Martin TM, Young DM. Prediction of the acute toxicity (96-h LC50) of organic compounds to the fathead minnow (*Pimephales promelas*) using a group contribution method. Chem Res Toxicol 2001;14:1378–85.
Mattei M, Hill M, Kontogeorgis GM, Gani R. A comprehensive framework for surfactant selection and design for emulsion based chemical product design. Fluid Phase Equilib 2014a;362:288–99.
Mattei M, Yunus NA, Kalakul S, Kontogeorgis GM, Woodley JM, Gernaey KV, et al. The virtual product-process design laboratory for structured chemical product design and analysis. Comput Aided Chem Eng 2014b;33:61–6.
Nannoolal Y, Rarey J, Ramjugernath D, Cordes W. Estimation of pure component properties: Part 1. Estimation of the normal boiling point of non-electrolyte organic compounds via group contributions and group interactions. Fluid Phase Equilib 2004;226:45–63.
Nannoolal Y, Rarey J, Ramjugernath D. Estimation of pure component properties: Part 2. Estimation of critical property data by group contribution. Fluid Phase Equilib 2007;252:1–27.
Ng LY, Chemmangattuvalappil NG, Ng DKS. Optimal chemical product design via fuzzy optimisation based inverse design techniques. Comput Aided Chem Eng 2014;33:325–9.
Garnier E, Bliard C, Nieddu M. The emergence of doubly green chemistry, a narrative approach. Eur Rev Ind Econ Policy 2012;4:1.
Ourique JE, Silva Telles A. Computer-aided molecular design with simulated annealing and molecular graphs. Comput Chem Eng 1998;22S:S615–8.
Panteli E, Saratsioti P, Stamatis H, Voutsas E. Solubilities of cinnamic acid esters in organic solvents. J Chem Eng Data 2010;55:745–9.
Papadopoulos AI, Stijepovic M, Linke P, Seferlis P, Voutetakis S. Molecular design of working fluid mixtures for organic Rankine cycles. Comput Aided Chem Eng 2013;32:289–94.

Prausnitz JW, Lichtenthaler RN, de Azevedo GE. Molecular thermodynamics of fluid phase equilibria. 3rd ed. New York: Prentice-Hall; 1998.
REACH. Regulation text, corrigendum and amendments; 2006, Available from: http://ec.europa.eu/enterprise/sectors/chemicals/reach/index_en.htm
Rice P, Teja AS. A generalized corresponding-states method for the prediction of surface tension of pure liquids and liquid mixtures. J Colloid Interface Sci 1982;86:158–63.
Riedel L. Eine neue universelle Dampfdruckformel Untersuchungen über eine Erweiterung des Theorems der übereinstimmenden Zustände. Teil I. Chem Ing Tec 1954;26:83–9.
Sinha M, Achenie LEK. CAMD in solvent mixture design. In: Achenie LEK, Gani R, Venkatasubramanian V, editors. Computer aided molecular design: theory and practice. Amsterdam: Elsevier; 2003. p. 261–87 [Chapter 11].
Solvason CC, Chemmangattuvalappil NG, Eden MR. A systematic method for integrating product attributes and molecular synthesis. Comput Chem Eng 2009;33(5):977–91.
Sova M. Antioxidant and antimicrobial activities of cinnamic acid derivatives. Mini-Rev Med Chem 2012;12:749–67.
Tamura M, Kurata M. On the viscosity of binary mixture of liquids. Bull Chem Soc Jpn 1952;25(1):32–8.
Vaidyanathan R, El-Halwagi M. Computer-aided synthesis of polymers and blends with target properties. Ind Eng Chem Res 1996;35:627–34.
Veith GD, Konasewich DE. Structure–activity correlations in studies of toxicity and bioconcentration with aquatic organisms. Windsor, ON: Great Lakes Research Advisory Board; 1975.
Venkatasubramanian V, Chan K, Caruthers JM. Computer-aided molecular design using genetic algorithms. Comput Chem Eng 1994;18:833–44.
VOC. VOC Solvents Emissions Directive (Directive 1999/13/EC) amended through article 13 of the Paints Directive (Directive 2004/42/EC); 2004.
Weis DC, Visco DP. Computer-aided molecular design using the signature molecular descriptor: application to solvent selection. Comput Chem Eng 2010;34:1018–29.
Yunus NA, Gernaey KV, Woodley J, Gani R. A systematic methodology for design of tailor-made blended products. Comput Chem Eng 2014;66:201–13.