

DEVELOPMENT OF A COMPUTER-AIDED PRODUCT DESIGN TOOL FOR SUSTAINABLE CHEMICAL PRODUCT DEVELOPMENT

M. TELES DOS SANTOS¹, V. GERBAUD¹

¹ Institut National Polytechnique de Toulouse (France)/CNRS, Laboratoire de Génie Chimique
E-mail para contato: moises.telesdossantos@ensiacet.fr

RESUMO – A computer aided product design (CAPD) tool was developed aiming to find bio-based molecules that match a set of desired properties. Molecules are represented by molecular graphs that are split into basic groups by automatic decomposition methods. These groups are further used for property predictions using Group Contributions methods by the property calculation component, a Dynamic-Link Library (DLL) that handles properties predictions for each candidate molecule generated by the search component. The search component uses a Genetic Algorithm to generate/modify molecular structures and search for molecules/mixtures that satisfy the desired properties. The capabilities of the tool are illustrated through a case study aiming to find chlorinated paraffin substitutes using levulinic acid as building block for alternative molecules.

1. INTRODUCTION

There is a social driver that boosts the search for improved products in the chemical industry. New products need to obey environmental, health and safety constraints in addition to usual product and process requirements. For finding alternative molecules for classical and new uses, the usual ‘trial and error’ approach seems inefficient unless high throughput screening is used. The combinatorial nature of the problem can lead to thousands or millions of potential solutions and their experimental tests can be time and economic resources consuming. Aiming to aid experiments to be focused in the most promising molecular structures, reverse engineering approaches, like Computer Aided Molecular Design (CAMD) have been proposed (Gani, 2004).

CAMD aims at finding molecules that satisfy a set of target properties defined in advance. CAMD relies upon four main concepts, namely, a molecular representation model, a set of property calculation models, a solving method and a performance criterion. Molecular fitness is evaluated thanks to property prediction models by comparing the values of estimated property and the target property. Then they are discriminated according to their performance and either improved, kept as is or rejected, with the help of the solving algorithm. This paper presents a computer aided molecular design tool (IBSS) and its tailoring for finding alternative bio-sourced molecules, in which a biomass based molecular fragment is fixed and simpler groups are used to generate molecules matching the desired properties. The IBSS tool was developed in the context of the French ANR CP2D 2009 project InBioSynSolv aiming at designing biosolvents.

2. COMPUTER-AIDED MOLECULAR DESIGN FEATURES

The main components of the software are described in Figure 1. It was developed using Model Driven Engineering concepts and is further detailed in Heintz *et al.* (2014).

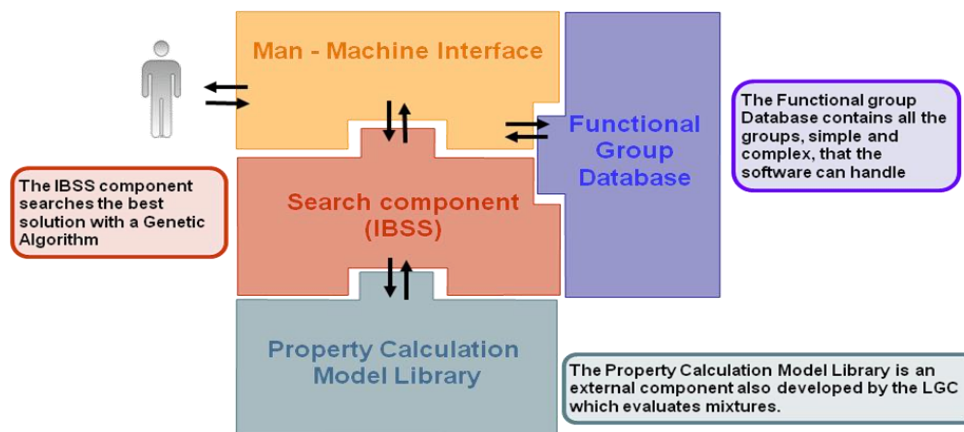


Figure 1 – Main components of the IBSS software tool.

2.1. Functional Group Database

A set of simple chemical fragments (e.g.: $-CO-$, $-OH$, CH_2-) are used to build molecules. In addition to these simple groups, larger molecular fragments can be also used. They can be tailored to cover biomass-based fragments, such as levulinic acid, or to generate molecules using a particular molecular fragment of interest, such as glycerol. These groups are stored and handled in a *XML* file. The computational representation of molecular fragments is the molecular graph. Using this representation, molecules can be generated, modified and their properties can be estimated using group-contribution based methods.

2.2. Property Calculation Model Library

This corresponds to a Dynamic-Link Library (DLL) written in Visual Basic that contains several property models based in the group-contribution concept. The decomposition of molecular graphs into molecular groups is done automatically by several algorithms written in Visual Basic.

2.2. Search Component

An initial population of molecules is generated. Then, a Genetic Algorithm written in C# is used to modify these molecular structures by using genetic operators adapted to cover the molecular design problem (deletion, insertion, crossover and mutation of chemical fragments). For each generated molecule, the actual property values are compared to the desired values and a performance (ranging from 0 to 1) is used as fitness evaluation. This overall performance is calculated using a Gaussian function and takes into account all concerned properties weighted by chosen values. The higher the

difference between the estimated and the desired value, the lower the performance. If the property is within target, performance is equal to 1. The search continues until a pre-defined number of generations is achieved. For further case studies and detailed features of the tool, the reader must check others works from the authors (Heintz *et al.*, 2012, 2014).

3. CASE STUDY: SUBSTITUTION OF CHLORINATED PARAFFINS

Chlorinated paraffins (CPs) are chlorinated linear chain alkanes with the general formula $C_nH_{2n+2-x}Cl_x$. These compounds 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). Due to their toxicity and potential carcinogenic properties, there are regulations and concerns about their use and risks (ECHA, 2009). Indeed, even its production process involves toxic substances (chlorine) and non-renewable raw materials (petroleum). This is a major drive to search for alternative molecules, with enhanced environmental, health and safety properties, maintaining the physical-chemical performance of CPs. Table 1 summarizes the physical properties to be satisfied by a potential substitute candidate.

Table 1 – Real and calculable properties with corresponding CAMD parameters.

Real Property	Calculable Property	Weight	Model*	Target	Tolerance	Performance at Tolerance
Liquid at usage temperature	Melting Point	1	HSKASG2012	< 10°C	10 °C	0.8
Liquid at usage temperature	Boiling Point	1	HSKASG2012	> 250°C	5 °C	0.8
Non flammability at high temperatures	Flash Point	1	HSKASG2012	> 232 °C	5 °C	0.8
Viscous	Viscosity	1	JR1987	> 37 cPoise @25°C	10	0.8
Low potential to bioaccumulation	Octanol/Water partition coefficient (logKow)	1	HSKASG2012	< 3	2	0.8

HSKASG2012 : Hukkerikar *et al.* (2012); JR1987: Joback and Reid (1987).

Building blocks: Levulinic acid (LA) has been evaluated as a platform chemical that can be produced cost effectively and in high yield from renewable feedstocks. LA and its derivatives have found use in many areas, such as lubricants, coatings and printing/inks (Bozell *et al.*, 2000). Seeking for more sustainable alternatives for chlorinated paraffins, the present methodology uses LA as the “fixed group” in the molecule. Figure 2 summarizes the molecular structures involved in the problem.

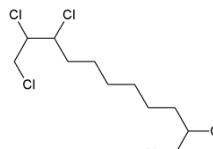
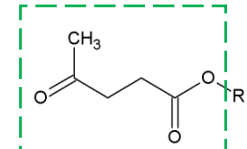
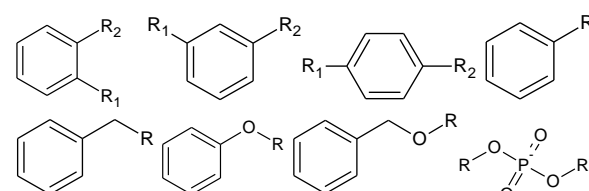
Molecule to be replaced	Biosourced group (platform to potential candidates)
 <p>Chlorinated paraffin example (environmental and health constraints)</p>	 <p>Levulinic Acid derivatives (renewable feedstocks)</p> <p>CH₃- -CH₂- >CH- >C< -C= CH₂= -O- -OH O= -COOH -COO- -CO- -COH -CH=CH- CH₂=CH- -Cl -Br -I -OCOO-</p> <p>(R: Basic Groups)</p>  <p>(R: Complex Groups)</p>

Figure 2 – A chlorinated paraffin example and building blocks used to generate alternative molecules.

Genetic Algorithm parameters: Number of generations (300), population size (100), elitism (10), crossover (P=10), mutation (P=50), insertion (P=5), deletion (P=5). Also, the formation of cyclic and aromatic compounds is not allowed. The maximum number of allowed groups (kmax) changes in different runs.

Set levulinic + basic groups: The maximum performance (0.9976) is achieved in 100 generations. The best 10 molecules of the 100th generation are given in Figure 3, along with a general schema for the present approach. Table 2 shows the predicted values for each molecule shown in Figure 3. The flash point could not be estimated due to property models limitations. There are missing contributions for the flash point for the group >CO (molecules 1 to 8 and 10), and for the group CHCO (molecule 9).

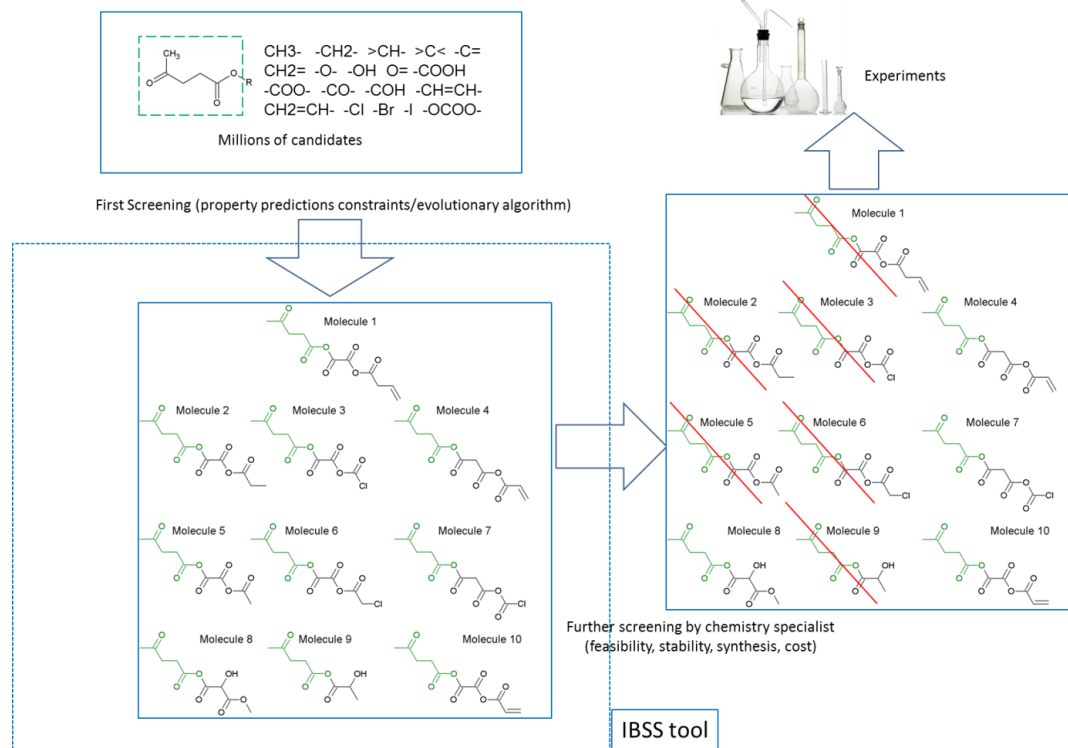


Figure 3 – Generated candidates for chlorinated paraffins substitution.

Table 2 – Predicted properties for the best 10 candidates for chlorinated paraffins substitution (100th generation, kmax = 5, levulinic acid + basic groups)

	Performance	Melting Point (°C)	Boiling Point (°C)	Viscosity (cP)	log (Kow)
molecule 1	0.9976	10.99	317.97	35.19	-0.72
molecule 2	0.9884	6.93	307.51	32.38	-0.96
molecule 3	0.9543	19.21	305.13	34.83	-1.11
molecule 4	0.9507	19.71	322.13	35.19	-0.83
molecule 5	0.9506	11.18	297.72	27.15	-1.45
molecule 6	0.9504	19.96	317.23	41.33	-0.99
molecule 7	0.9504	19.96	317.23	41.33	-0.99
molecule 8	0.9411	20.97	306.97	128.89	-1.65
molecule 9	0.9379	21.31	279.57	54.15	-1.07
molecule 10	0.9302	18.96	310.27	29.60	-0.96

Also, all molecules have predicted melting points greater than the target (exception for the second molecule). However, the deviations from the target are within the reported average absolute error of the model (17.65 K), and these molecules can be retained for further experimental analysis of their melting point. As the performance does not take into account the easiness of synthesis,

feasibility or stability, the potential candidates must be further screening taking into account these issues (Figure 3). Indeed, as the group –CO-CO- is known to be unstable, 4 molecules are retained for further tests.

Set levulinic + basic groups + complex groups: Figure 4 shows the influence of the size of the molecule (kmax) on the GA performance. Using only 3 or 4 building blocks, no solution with high performance is obtained. As the number of allowed building blocks increases (5, 6 and 7), higher performance solutions are obtained. Also, with 7 possible building blocks, the algorithm is capable of finding optimal solutions (performance = 1) with fewer generations (40).

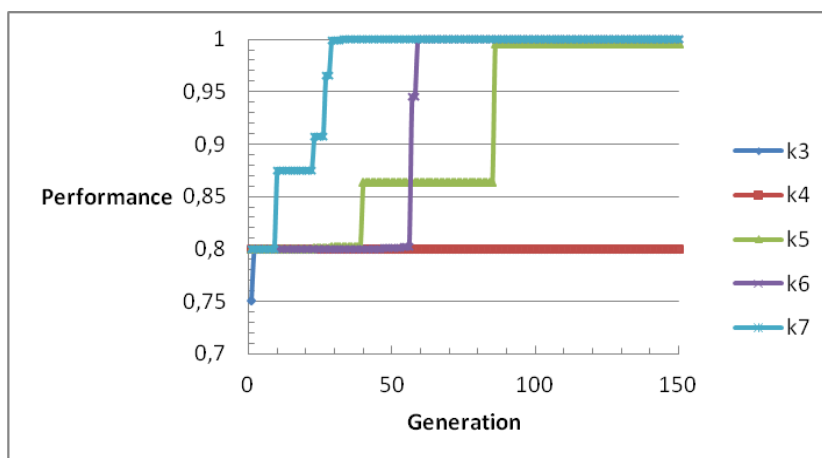


Figure 4 – Influence of the size of the molecules in the Genetic Algorithm performance.

Figure 5 shows the top 10 molecules arising from the 40th generation with kmax = 7. The corresponding properties of the best 5 molecules are shown in Table 3. It was not possible to compute the viscosity of such structures, because the Joback and Reid method does not have the group phosphate in its set of groups. The method of Hukkerikar *et al.* (2012) does not have the contribution of the group *aC-PO4* to the boiling temperature; therefore, the boiling temperatures values shown on Table 3 are only first approximations that do not take into account the *aC-PO4* group.

Set basic groups + complex groups: As an additional example of the methodology, a run without the levulinic acid was performed. The objective is to evaluate the capability of the computational approach to find known current alternatives for CPs. Indeed, the 3th molecule of the 100th generation (kmax = 4), has great similarities with known CPs alternatives (Figure 6). These examples show that present methodology is able to converge to good candidates, exploring the combinatorial problem in a pre-experimental step. The GA was also able to determine many high-fitness alternative molecules. However, some of the near-optimal solutions show unstable structures and/or high deviations concerning only one property. As the GA is allowed to run up to

300 generations, and all intermediate structures are saved, the whole set of generated molecules and corresponding properties can be further analyzed.

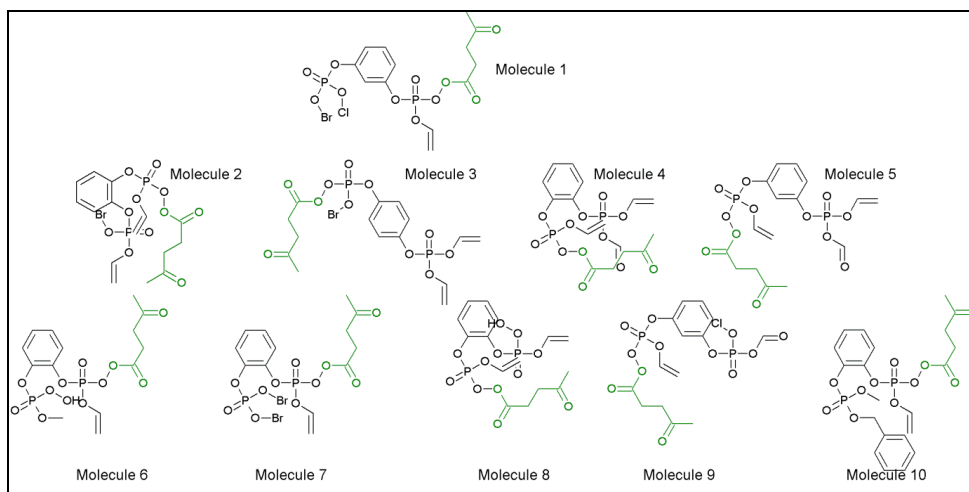


Figure 5 – 10 molecules with higher performance for chlorinated paraffins substitution. Building blocks: levulinic + basic groups + complex groups. 40th generation, kmax = 7.

Table 3 – Predicted properties for the best 5 candidates for chlorinated paraffins substitution (40th generation, kmax = 7, levulinic + basic groups + complex groups).

	Performance	Melting Point (°C)	Boiling Point (°C)	Flash Point (°C)	log (Kow)
molecule 1	1.0000	1.65	358.67	245.80	0.31
molecule 2	1.0000	-6.81	360.24	234.15	0.30
molecule 3	1.0000	10.15	362.01	247.08	0.36
molecule 4	0.9986	11.58	361.60	233.83	-0.97
molecule 5	0.9982	11.79	362.59	254.61	-0.82

4. CONCLUSIONS

Although the limitations due to property predictions models, the CAMD approach offers a large set of potential solutions ranked by performance, allowing the experiments to be focused in the most promised structures. As in all CAMD approaches, the final step is a further screening regarding stability, feasibility of synthesis and cost. The algorithms developed to automatically identify molecular groups and the use of the state of the art of group-contribution methods offer an alternative to deal with the large combinatorial problem faced in the early stages of molecular design.

5. ACKNOWLEDGEMENT

This scientific work was supported by the French national research agency (InBioSynSolv ANR-CP2D-2009-08).

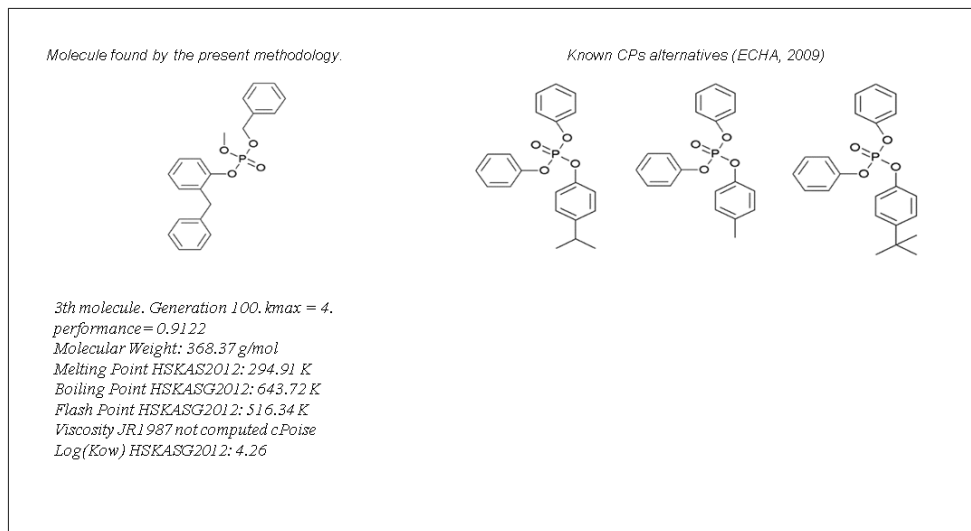


Figure 6 – Known alternatives for chlorinated paraffins and a similar structure found by the Computer-Aided Molecular Design approach.

6. REFERENCES

- BAYEN, S.; OBBARD, J.P.; THOMAS, G.O. Chlorinated paraffins: A review of analysis and environmental occurrence. *Environ. Int.*, v. 32, p. 915–929, 2006.
- BOZELL, J.J.; MOENS, L.; ELLIOTT, D.C.; WANG, Y.; NEUENSWANDER, G.G.; FITZPATRICK, S.W.; BILSKI, R.J.; JARNEFELD, J.L. Production of levulinic acid and use as a platform chemical for derived products. *Resour. Conserv. Recycl.*, v. 28, p. 227–239, 2000.
- ECHA: European Chemicals Agency, 2009. Background document for alkanes, C10-13, chloro (SCCPs).
- GANI, R. Chemical product design: challenges and opportunities. *Comput. Chem. Eng.*, v. 28, p. 2441–2457, 2004.
- HEINTZ, J.; BELAUD, J-P.; GERBAUD, V. Chemical enterprise model and decision-making framework for sustainable chemical product design. *Comput. Ind.*, v. 65, p. 505–520, 2014.
- HEINTZ, J.; TOUCHE, I.; TELES DOS SANTOS, M.; GERBAUD, V. An integrated framework for product formulation by computer aided mixture design. *Computer Aided Chemical Engineering*, v. 30, p. 702–706, 2012.
- HUKKERIKAR, A.S.; SARUP, B.; KATE, A.T.; ABILDSKOV, J.; SIN, G.; GANI, R. Group-contribution+ (GC+) based estimation of properties of pure components: Improved property estimation and uncertainty analysis. *Fluid Phase Equilib.*, v. 321, p. 25–43, 2012.
- JOBACK, R.; REID, R.C. Estimation of Pure-Component Properties from Group-Contributions. *Chem. Eng. Comm.*, v. 57, p. 233–243, 1987.