Kevin Adrián Rodríguez Ruiz

Including toxicity and market availability in a Computer-Aided Molecular Design methodology

MASTER'S FINAL WORK
To be presented for the Degree of
Master of Engineering in Systems and Computing

Facultad de Ingeniería de Sistemas e Industrial,
Universidad Nacional de Colombia

Bogotá, July 2019

MASTER'S FINAL WORK IN COMPUTER AIDED MOLECULAR DESIGN

Advisor:	Juan Carlo Serrato, Ph.D. Facultad de Ingeniería Química y Ambiental, Universidad Nacional de Colombia, Bogotá, Colombia
Co-advisor:	Jonatan Gómez Perdomo, Ph.D. Facultad de Ingeniería de Sistemas e Industrial, Universidad Nacional de Colombia, Bogotá, Colombia

To my father, my role model.
I'm sorry you couldn't see me finishing this stage of my life.

To my brother David, who is always in in my mind

To my mother, the strongest woman I know

To my aunt Nelly, a very appreciated member of the family.

To my friends, to support me in the good times and the hard times

And to those who value my actions more than my words

Bogotá, July 11, 2019
Kevin Adrián Rodríguez Ruiz

Abstract

Advances in the understanding of natural phenomena and the exponential increase of computing power over the last years have made possible the solution of chemical product design problems using computational approaches. In this work, a Computer Aided Molecular Design (CAMD) methodology is proposed and implemented for the design of environment-friendly solvents in liquid-liquid extraction. As the proposed methodology aims to solve a problem of chemical industry, to ensure that the designed solvents can be easily acquired or synthetized, market availability criteria are included.

The proposed CAMD methodology formulates and solves a multi-objective optimization problem where the decision variables are molecules represented as chemical graphs. In the definition of this problem, a first objective is the maximization of solvent power and a second objective is the minimization of environmental impact. Market availability is included in the methodology as one constraint of the optimization problem.

In optimization, molecules require specific encodings and the usage of flexible methods. Hence, in the methodology proposed, the evolutionary algorithm HAEA is selected to perform optimization as this algorithm allows flexibility in the representation of individuals and the inclusion of custom genetic operators. The original HAEA is intended to solve single-objective optimization problems, then this work proposes and implements a multi-objective version of HAEA (MOHAEA) for the solution of the optimization problem contained in the CAMD methodology. In MOHAEA, Pareto optimality and the NSGA-II crowding-distance are used to evaluate solutions and guide the evolution. In addition, a strategy for the handling of constraints based on Pareto front punishment is proposed in this new algorithm.

The methodology presented in this document is an extension of the CAMD methodology presented by Serrato in 2009. The Serrato's work is the starting point of this work and many of the methods persist in the methodology proposed. Serrato addresses the chemical product design problem of designing optimal solvents for liquid-liquid extraction using a single-objective optimization approach. The study case in both works is the design of optimal solvents for the separation of lactic acid from an aqueous solution and the major improvement of the new methodology proposed is the design of solvents with similar solvent power, a significant reduction of environmental impact and a market availability greater than 80%.

Contents

Tables

Figures

1 Introduction

From the early decades of chemical industry, discovering compounds with new features, inclusion of substances with improved performance and releasing valuable products to the market have been common objectives in a great part of the research work in this industry. The process of searching a chemical product or blend of chemical products that exhibits a desirable or specific behaviour is named *Chemical Product Design* [1].

In the results of research in chemical industry along the twentieth century, there is a large number of products that, once released, improved the quality of life of people. Some of the most relevant examples are petroleum based products such as plastics, solvents or detergents; or products used in pharmaceutical industry as active compounds or as additives for enhancing performance of drugs. In that time, chemical industry gained more and more importance in the course of human activities, no matter that methods used in chemical product design remained basically untouched. There was no significant evolution, since most of the research was conducted via traditional experimental approaches.

These experimental approaches consist mostly on trial-error work in laboratory and this lead to inevitable limitations [2]. Namely, the limitations are the amount of time, financial resources and compounds availability. In addition, the regular set of chemical compounds that can be considered and evaluated is limited, then the so-called *chemical design space* is very narrowed compared with the theoretical design space.

Nowadays, the market and the chemical industry demand for more specialised chemical products, mainly high value-added products whose performance is more important than the composition [3]. To overcome the limitations of traditional experimental approaches and meet the new requirements of industry, new approaches in chemical product design have emerged. In line with this reality, the growth of computational power and the easy access to powerful computers have made possible the solution of product design problems using Computer-Aided Molecular Design (CAMD).

CAMD is a very promising computational approach aimed to solve problems of chemical product design making use of methods for prediction of physical-chemical properties combined with efficient algorithms to design, evaluate and select optimal molecules. In CAMD, product design tends to be represented as an optimization problem where the objective is the desired property and the decision variable is the composition of the chemical product. The major benefit of using

CAMD compared with the traditional approach is the reduction of experimental work and the possibility of explore a larger chemical design space.

In the context of the discovery of more specialised products, designing chemical products meeting not only the desired specifications but also environmental performance requirements have become a trend lately. In CAMD, a wide search space can lead to the design of chemical products with dangerous toxicity indicators. If a product causes negative affectations to living beings or the environment, it will also lead to produce difficulties at the time of its introducing to the market or its inclusion in a large-scale process. Designing environment-friendly products in CAMD requires methodologies including more than one desired features and multi-objective optimization algorithms.

Professor Juan Carlos Serrato[1] presents a doctoral dissertation entitled "*Computational design of extraction agents for the separation of organic compounds in aqueous streams, application to lactic acid*" [4] in 2009 at the Universidad Nacional de Colombia. In that work, Serrato proposes a CAMD methodology aimed to design optimal solvents for the extraction of acetic acid and lactic acid from aqueous streams.

The methodology of Serrato addresses the problem of designing compounds with optimum values for selectivity and distribution coefficient. Both properties are very important in liquid-liquid extraction operations. Serrato formulated an optimization problem where the objective function to maximize is the product between selectivity and distribution coefficient for the designed compounds. The result of that work is a java program capable of designing molecules with favourable values for selectivity and distribution coefficient, however none of the molecules reported by Serrato to be optimal for the separation of lactic acid are available in catalogues of the major chemical product suppliers. In addition, compounds with similar structure to the designed ones, but available in the market, hold high levels of toxicity.

In line with the environmental considerations mentioned above, this work presents a new Computer-Aided Molecular Design methodology, based on the previous one proposed by Serrato, but including toxicity and market availability and address the problem using a multi-objective optimization approach. This work aims to achieve the following goals:

1. Reformulate the optimization problem contained in the CAMD proposed by Serrato in order to consider market availability and toxicity.
2. Establish and implement methods to determine the market availability and the toxicity of a given chemical compound.
3. Propose and implement new genetic operators and molecular construction methods in the CAMD methodology to perform a better exploration of the chemical search space.
4. Implement a multi-objective optimization algorithm to solve the optimization problem of the CAMD methodology.

[1] Chemical Engineer Ph. D., associate professor in the Department of Environmental and Chemical Engineering, Universidad Nacional de Colombia

5. Develop a new version of the CAMD program with the new methodology.
6. Evaluate the performance of the new CAMD methodology compared to the previous methodology proposed by Serrato.

The second chapter is a conceptual review of the foundations of CAMD. The chapter introduces the Chemical Product Design problem, the Computer-Aided Molecular Design approach, methods for estimating properties in CAMD, techniques used in the solution of CAMD problems, and the importance of CAMD in the design of environment-friendly chemical products.

The third chapter contains the elements comprising the proposed CAMD methodology. The multi-objective optimization problem is formulated. The properties used to evaluate the separation power and the environment performance of a solvent in liquid-liquid extraction, as well as the available mathematical models to predict those properties, are described. The strategy to assign the market availability of a compound, including the queried databases is described. And finally, the multi-objective optimization algorithm used to solve the CAMD problem is explained.

The fourth chapter discusses the results of the new CAMD methodology. In the chapter, the performance of the new methodology is compared with the Serrato's methodology. The impact of including market availability and toxicity in the separation problem is analysed. And the performance of the new genetic operators is evaluated.

Finally, the fifth chapter presents the conclusions and future work. The contributions of this work are summarized and new possible developments for this work are proposed.

2 Theoretical Background

2.1 CHEMICAL PRODUCT DESIGN

A *chemical product* is defined by Cisternas [5] as an arrangement of chemical substances which are manufactured for one or more purposes. This definition implies that a chemical product requires at least one purpose, then not all chemical substance or mixture of substances can be considered as a chemical product. A common classification of chemical products is as follows [3]:

- *Commodity chemicals*: the goal is to produce these products at the minimal cost, these are the most widely produced in the world. Ammonia and acetic acid are in this classification.
- *Molecular products*: such as pharmaceutical products, the goal for these products is achieving a fast discovery and a fast introduction to the market.
- *Performance products*: for these products, the goal is to develop products with certain functionality, regardless the chemical composition. Lubricants are an example of this category.

Now, *chemical product design* can be defined as the entire process in which a chemical product is determined. Moggridge and Cussler [6] proposed a four steps scheme for describing this process. Initially, the specifications that the product must fulfil are identified according to the customer requirements or a particular problem (*Needs* step). Next, different pictures of how this product could be are posed (*Ideas* step). Then, a mechanism for selecting the best product is designed according to the given ideas (*Selection* step). Finally, the design of the manufacturing process for the selected product starts (*Manufacturing* step). When the desired product requirements cannot be met by a single component, an optimal mixture of chemicals must be considered [7].

The key in the solution of a chemical product design problem is the selection of the best "idea for product" (*Selection* step), hence the attention in the study of chemical product design has been focused on this aspect. One difficulty of adopting this perspective is that a chemical product design problem becomes the inverse of a property prediction problem. Product design identifies chemical product candidates that match best with the desirable values for a defined set of properties [8], and as many reverse problems, the solution cannot be reached using direct methods. Traditional *bottom-up* approaches have been used in problems of chemical product design. These approaches consist of heuristics, experimental studies and expert knowledge, involving a large effort of trial-error and expensive experimentation [9]. Multiple candidates are evaluated in the laboratory in order to check if those meet the desired product properties. L.Y. Ng [10] remarks

that as the chemical industry moves towards the manufacturing of more value added chemical products and more complex requirements. For these products, the bottom-up approaches are unsatisfactory in terms of effectiveness and resources investment.

The last years, Computer-Aided Molecular Design (CAMD) has gained relevance as a future tool for solving problems of chemical product. CAMD consists in the use of computational methods for predicting, estimating and designing molecules starting from a set of predefined target properties in order to reduce the experimental effort and the usage of biased heuristics.

2.2 COMPUTER AIDED MOLECULAR DESIGN (CAMD)

As defined by Austin [2], Computer Aided Molecular Design (CAMD) is a computational approach for the solution of chemical product design problems, the objective is design good or optimal molecules structures combining molecular modelling techniques, thermodynamics, and numerical optimization. With CAMD techniques is possible to identify molecules with certain properties of interest without the need of performing the arduous task of synthesizing and testing them experimentally [7].

In the traditional *bottom-up* approach for solving chemical product design problems, the design process consists in identifying a set of molecules and then checking if those meet the final product requirements. In contrast, CAMD can be considered as a *top-down* approach, because as Gani describes [11], CAMD starts with the definition of the properties the chemical product needs to fulfil and then the CAMD methods searches for the molecules whose properties meet the product requirements. The top-down approach makes possible not only the efficiently identification of the desired product, but it explores a larger section of the chemical design space [2].

A CAMD methodology for the solution of chemical product design problems usually consists of the steps presented in the process flow of Figure 1 [12]. Such steps were identified by Gani and Harper [13] and consist of:

- **Pre-design**: The requirements of the desired product are stablished, this step includes the selection of the desired properties that must be optimal as well as other features that the desired product must meet, including boundaries for other properties and concerns related to the chemical structure of designed optimal product.
- **Design**: The definitions from the previous step are transformed into objectives and constraints. Properties prediction models, molecular modelling methods and optimization algorithms are combined in order to design the feasible chemical compounds most suitable for being part of the resulting candidate compounds list.
- **Post-design**: A detailed analysis is performed on the candidate compounds generated in the design step and the concerns related to the manufacturing or acquisition of the chemical compounds (product) are posed.

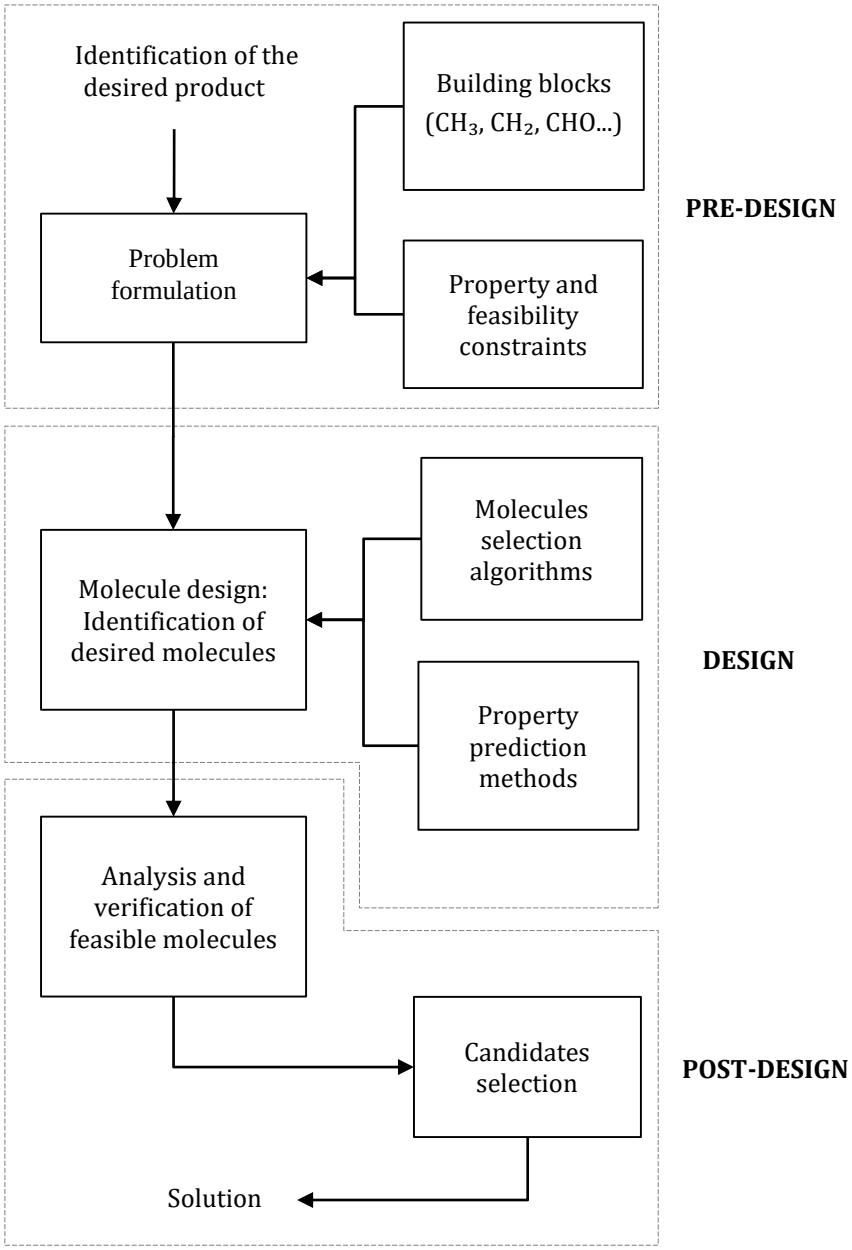


Figure 1. Stages in a CAMD methodology. Adapted from the Roughton [14]

Regarding the design step in Figure 1, the unavailability of properties prediction methods and selection algorithms for candidate molecules is the reason why the appearance of CAMD techniques emerged just a few decades ago. The implementation of the first CAMD methodologies in the 1990's was possible thanks to the development of group contribution methods capable of predicting pure component and mixture properties as well as the emergence of efficient optimization algorithms able to solve combinatorial problems. The initial works in CAMD consisted in the identification of optimal solvents for separation operations using mathematical programming [15] and generate-and-search [11] algorithms for the selection of the best compounds. In those works, prediction of properties for the compounds evaluated is done by using group contribution methods.

2.3 PREDICTION METHODS FOR PROPERTIES IN CAMD

CAMD can be seen as the problem of predicting chemical structures from properties. However, today there is not a known method capable to immediately relate certain chemical structure and its performance for certain application [16]. In CAMD, a large number of structures are systematically evaluated and an important computational effort is required, hence efficient methods for the quantification of properties are required for this process.

Semi-empirical quantitative structure property relationships (QSPR) are the preferred methods for addressing the prediction of properties. QSPR methods consider a chemical structure as a set of sub-structures comprising bonds and atoms, and use these elements to compute the desired properties.

2.3.1 Group-contribution methods

The group contribution (GC) approach is the most popular in CAMD for predicting properties with QSPR. The estimation of a property using GC consists in listing the occurrences of predefined contributions groups in the molecule structure. These methods are simple to apply and provide a quick and accurate prediction for many properties without requiring significant computational resources [17].

The first GC methods of 1980's assume that contribution groups are independent entities and proximity effects among them are not considered. The result is lack of reliability in predictions as the complexity of the molecules increases and no possibilities for distinguishing isomers.

An improved GC method is presented by Constantinou and Gani (1994) [18]. This method includes molecular groups of first and second order. The first order groups address the basic group contributions for properties estimation, while the second order groups address the differentiation among isomers and in some extent interactions among the first order groups. This approach bring improvements in terms of accuracy, however the application ranges are still restricted [18].

Years later Marrero and Gani (2001) [19] extend the above method including third order groups into the model. The new groups address the interactions among the functional groups for which the first and second order flawed in predicting [2]. This method, referred as the GC+ method, brings a significant improvement of accuracy and applicability. Today, this method is the most widely used GC method in CAMD. Estimation of properties using the GC+ is achieved using the following equation:

$$f(X) = \sum_i N_i C_i + w \sum_j N_j C_j + z \sum_k N_k C_k \qquad \textbf{Equation } \textit{1}$$

Where f(X) is a function for the target property X, while w and z are binary coefficients set according to the level of estimation, using first-order and second-order groups, respectively; N_i, N_j and N_k are the numbers of occurrences for first, second, and third order groups, respectively. C_i , C_j and C_k are the contributions of the first, second, and third order groups, respectively. The values for C_i , C_j and C_k can be found in tables [19].

2.3.2 Topological indices

The topological indices (TI) approach for predicting properties with QSPR is based on the concept of chemical graph. According to the chemical graph theory [20], atoms and bonds which constitute a molecule can be represented as the set of nodes and edges of a molecular graph. In TI, an index corresponds to a descriptor for the interactions among different atoms or molecular groups inside a molecular graph.

The properties of a molecule can be computed through its index. The most known topological indices [10] are the Wiener indices [21], the Randic's molecular connectivity index [22] and the Kier's shape indices [23].

TI are very useful in problems where exists knowledge about some structural features that the desired product must have. In addition, TI is recommended in problems where is important to differentiate among very similar structures, like isomers, GC methods do not work well with most of the types of isometry. Properties prediction in pharmacological CAMD problems have been addressed successfully with TI.

The main drawback of TI, compared with GC, is that its incorporation in CAMD methodologies is more difficult. Besides, the applicability of TI is restricted to certain class of chemicals, then they consider only a part of the chemical search space [2].

2.3.3 Signature descriptors

Signature descriptors (SD) emerged as the need of methods that make use of the advantages of group contribution methods and topological indices without requiring sacrifice computational performance [24]. While GC methods divide a molecule into the main subgroups of atoms and TI methods are based on chemical graphs, SD conceive a molecule structure as a chemical graph [25], but it stores and uses all the structural and connectivity information for each atom of the molecule.

In SD, an atom signature is the representation of its extended valences to a predefined height and the signature of a molecule corresponds to the linear combination of its atomic signatures [10].

2.4 Solving the CAMD problem

A CAMD methodology must perform in such way that it could evaluate a large number of structures without a great computational effort. Each compound evaluated must meet not only the properties of the desired product, but it must match with the molecule feasibility criteria, according with the rules of chemical compounds (such as the octet rule). Currently, numeric optimization is the main method for reaching the desired problems in CAMD, however there are more approaches.

2.4.1 Enumeration approaches

The first methodologies proposed for the solution of CAMD problems are based on the enumeration approach, also known as the *generate-and-test* approach. These methods are used in cases where the computation of the desired properties required little computational effort. Enumeration consists in the generation of a large number of feasible candidate molecules followed by the evaluation of the properties of interest for each molecule [2]. This method uses a short chemical design space and even when several techniques has been developed for reducing systematically the number of candidate molecules such as the developed by Harper and Gani [16], it is strongly not recommended for problems where the chemical design space is large or the desired product is expected to be a molecule whose properties require a relatively high computational effort.

2.4.2 Mathematical optimization

Mathematical optimization methods emerge to solve CAMD problems of greater complexity. The first formulation of a CAMD problem as an optimization is made by Odele and Macchietto [9]. Their problem is formulated as a mixed-integer nonlinear program (MINLP). The main problem with mathematical optimization is the bad performance in non-convex problems, in such cases reaching globally optimal solutions is not guaranteed.

Maranas [26] proposed a method where the properties methods are linearized to allow the solution of the optimization problem using mixed-integer linear program (MILP) techniques. An advantage of this approach is that in a CAMD problem formulated in terms of MILP the global optimal solution is guaranteed.

The available techniques for solving optimization problems can be classified into stochastic optimization techniques and deterministic optimization techniques [10]. In stochastic techniques, solution is reached using random choices for defining the search direction in each iteration of the method. In deterministic techniques, there are clear procedures to define the search direction and the size of the steps. Afterwards, the method advances in those directions along the iterations.

2.4.3 Metaheuristics

In many real-world problems, such as CAMD problems, reaching the global optimal solution is not possible by mathematical optimization as those problems are too high-dimensional or non-linear. Under this scenario, heuristic search techniques may be used to obtain quickly a good solution. The purpose of heuristics algorithms is to produce a good solution, but with low probabilities of being the optimal solution [27]. In addition, these algorithms can be easily considered as inconsistent and biased.

Metaheuristics algorithms has been designed to overcome the problems of mathematical optimization and heuristics. As Glover [28] has stated, these algorithms consists of strategies for guiding and modifying heuristics in order to construct solutions beyond those that normally result in local optimality. As a result, metaheuristics have gained a lot of popularity due to its flexibility, efficiency and

ease in the implementation. Many type of problems, including CAMD, can be solved using these algorithms.

2.4.3.1 Genetic Algorithms

Genetic algorithms (GAs) are optimization methods based on the idea of natural selection. The concept of GA, introduced in computing by Holland in 1975, is a process where a population consisting of "chromosomes" evolves successively to a new population by the action of genetic (inspired) operators that mimic "natural selection" [29]. Many methods have been proposed around this idea.

In GAs an individual is a possible solution for the problem, and it is represented in a "genetic" form named chromosome. During every generation of the evaluative process, there is a collection of individuals, and according to the proximity of each individual to the target, its fitness is assigned. The value of the fitness guides the selection process in order to allow the best individuals (better fitness) to survive while the others disappear.

In a genetic algorithm [10], a group of genetic operators are defined in order to transform the survivors of a population (selection process) and produce the members of the next generation. Crossover and Mutation have been the reference point for genetic operators. In crossover, two offspring chromosomes are created by exchanging contiguous fragments from the chromosomes of two individual parents. In the case of mutation, this operator modifies one or more units of a chromosome of one individual. The evolution process is repeated generation by generations until an acceptable solution is obtained according to certain criteria or until the algorithm reaches a predefined number of generations.

In CAMD, the work of Venkatasubramanian [30] is the first in include GAs into a CAMD problem, using the design of a polymer as case study. Van Dyk and Nieuwoudt [31] proposes an encoding for the molecules based on the UNIFAC groups. QSPR methods usually are combined with genetic algorithms.

2.4.3.2 Simulated Annealing

Simulated Annealing (SA) is based on the analogy between problem optimization and statistical physics. It consists on the random modification of an initial molecule. The transformation is aimed to transform the molecule into a new one with better performance. The performance is based on certain criteria stablished at the beginning of the problem. If the modification is successfully, the new molecule is retained for further modifications.

Compared with other metaheuristic methods, SA has proved to increase the possibility for producing solutions near the global optimal, however the nature of this method make it difficult its application to problems with more than one objective.

Simulated annealing in CAMD comes from the work of Ourique and Telles [32]. They applied SA in the design of refrigerants for utilization in heat pumps, and the selection of the best solvent in the separation of n-butanol from aqueous solution.

2.4.3.3 Tabu search

Tabu Search (TS) is an approach for solving combinatorial optimization problems. It starts proposing a pool of initial solutions. A set of operators are defined to alter slightly the solutions and produce a new pool. The procedure is repeated as long as the molecules generated do not appear in a tabu list [2]. The tabu list is the main feature of these approach, it consists of a list of previous solutions ("memory") that helps to guide the search process. The advantages of TS over other metaheuristic methods is that it does not get easily trapped upon local optima (good exploration) and is capable to identify near-optimal solutions (good exploitation).

TS is introduced by first time in CAMD problems by Lin and Chavali [33] for the design of transition-metal catalysts. In CAMD, TS algorithms are particularly useful due to the tabu list can contain: number of occurrences, unfeasible molecules or low fitness value ranked molecules.

2.4.4 Decomposition methods

Decomposition methods are employed when an optimization problem is very complex, and the most convenient approach to reduce difficulty is by dividing it into a series of optimization subproblems, each one addressing different elements for the solution, the subproblems share their results and are solved iteratively until a satisfying solution for the overall problem is reached. In CAMD, there are three particular types of problem where decomposition methods have been successfully applied.

In single molecule design, where constraints are very tight, decomposition methods have been used initially for the generation of several scenarios focused on different design sub-spaces [16]. Then start the search for the solution in the subspaces, one by one.

In the case of compounds mixture design problems, where the desired product is not a single molecule but a mixture of them, the complexity lies in the inclusion of the variables that affect mixture properties, mainly the composition. Besides, a high number of molecule-molecule interactions produces a bigger computational effort for properties prediction [7]. Decomposition methods for this problem consist in decoupling the mixture design problem into several single-component CAMD problems [34]. The solution of every problem addresses to find a candidate component for the mixture solution, and the information gained with the solution of the sub-problems is used for producing a new series of sub-problems, and so on.

Integrated product/process design problems are also often decomposed. These problems consist in defining not a set of target properties for the desired compound in the optimization, but process variables considering the inclusion of the designed compound into a complete chemical process. The difficulty of considering a process variable lies in the complexity of the objective function due to the inclusion of property prediction methods and methods for modelling the process (material balances, conditions, equipment, etc.). Despite of difficulties, decompo-

sition methods have been successfully used in these problems. A two-stage approach is proposes by Ede [35] consisting in the optimization of the process to determine the optimal properties of the desired chemical product, then identification of the product is done by CAMD techniques. Finally, the process is updated with the properties of the closest molecule found by CAMD and the process is re-optimized. Other approaches [2] consist in finding the optimal chemical product and then optimizing the process, or optimize both process/product at the same time.

2.5 CAMD FOR ENVIRONMENTALLY FRIENDLY CHEMICAL PRODUCTS

In chemical product design, the objective is the identification of compounds that accomplishes (or gets close to) the desired product requirements and most of the CAMD methodologies search through an abstract chemical design space, looking for the compounds whose properties suit best to the desired product properties. Nowadays, industry is looking for products and manufacturing environment-friendly processes [7] and CAMD is looking in that direction by adapting existing methodologies to make possible the design of chemical products with the desired properties and environment-friendly too.

According to the mentioned above, CAMD problems needs to be seen not as the problem of optimizing one desirable feature, but the problem of optimize the physical-chemical properties of the desired product and the properties addressing the reduction of the environmental impact of the product.

The first publication of a CAMD methodology considering environment-friendly features in the product is the design of environmentally safe refrigerants done by Duvedi (1996) [36]. That work uses mixed integer nonlinear programming (MINLP) for finding solution compounds and includes the environmental criteria as constraints in the formulation of the problem (ε-constraint optimization). Later, in the work of Pistikopoulos and Stefanis (1998) [37], a three-step process is developed for designing solvents by minimising environmental impacts. These steps presented involve:

- The identification of agent-based operations within the process of interest and specification of performance constraints.
- At the separation task level, the determination of a list of candidate solvents satisfying processing and environmental constraints.
- At the process level, the selection of an optimal solvent based on global plant-wide process and environmental constraints.

Buxton et al. (1999) develops a methodology for optimal solvent blends design with reduced environmental impact. Other works published after the mentioned correspond to the inclusion of known environmental criteria in different CAMD problems [38], including the design of metal catalysts and crystallization solvents.

An important contribution to this matter is done by Hukkerikar et al. (2012) [39]. They develop a model based on GC+ to provide reliable estimations of 22 environment-related properties of organic chemicals. This method eases the inclusion

of a more complete set of environmental properties into CAMD problems. A combination between process/product design including economic and environmental criteria is used by Ng et al. (2013) [40] for the synthesis of an integrated biorefinery.

In a recent work, Ooi et al. (2019) [41] aims to design solvents for the extraction of oil from Palm pressed fibre. In that work, solvents are designed by optimizing a function consisting of the weighting and the addition of nine properties, five of them environment-related properties.

Addressing CAMD as the problem of optimizing two objectives can lead to a contradictory behavior and manufacturing companies must evaluate the trade-off between product performance and environment impact. To handle this problem, *multi-objective optimization* approach has become an important tool in CAMD for the design of chemical products with more than one optimal feature. Below, several strategies used in CAMD to solve the multi-feature problem are described.

2.5.1 Weighted sum methods

Weighted sum methods are the most used approach in the solution of multi-objective problems in CAMD and in general. These methods consist in transforming the multiple objectives into a scalar objective function containing the sum of all the contributions made by each objective. The contribution of each objective consists in the product of the objective function by weighting factor. Mathematically it is expressed in Equation 2 [42]:

$$A^{\text{weighted sum}} = b_1 A_1 + b_2 A_2 + \cdots + b_m A_m \qquad \textbf{Equation 2}$$

Where $A^{\text{weighted sum}}$ is the overall objective function. b_m is the weighting factor of the individual objective function A_m.

The major drawback of these methods is that, in the assignment of appropriated weighting factors for each objective, there is no generally accepted criteria beyond the preferences of the decisions maker. As a result, these methods tend to be biased [43].

2.5.2 Bi-level Optimization

This approach, for multi-objective problems, consists in ordering the objectives of the problem into a hierarchy. The solution for the main problem is reached while its sub-problems are solved in hierarchical order. In this method, the decision maker must categorize the objectives into upper-level objectives and lower-level objectives [10].

Other type of by-level optimization is the called ε-constraint optimization. In this method, one objective is defined as the one to be optimized while the functions of the other objectives are converted into constraints by setting an upper bound to each of them.

2.5.3 *A posteriori* methods

This approach for multi-objective optimization can be considered as the less biased due to there is no interaction of the decision maker until the end. These methods do not provide a single solution for a problem but a set of optimal solutions represented as a Pareto optimal front [44].

In a problem of n objectives represented by the set of functions $F = \{F_1, F_2, \dots, F_n\}$, from a set of solutions, the selection of the solutions that conform the Pareto optimal front is performed according to the next definition:

> Pareto Optimal: A point $x^* \in X$ is Pareto optimal if there is no other point, $x \in X$, such that $F(x) \leq F(x^*)$, and $F_i(x) < F_i(x^*)$ for at least one objective function. The notation used to indicate that a solution x^* dominates a solution x is $x^* \prec x$.

None of the solutions in the optimal front can be considered better or worse than the others and the decision maker imposes preferences over these solutions. The generation of solutions for a multi-objective problem can be computational expensive depending of the size of the Pareto optimal front generated.

3 Methodology

The aim of this work is the design of a CAMD methodology for the design of solvents with a good performance in the separation of liquid mixtures, a reduced impact on environment and available in catalogues of chemicals vendors. This chapter presents the decision process involved in the selection and implementation of the methods used to achieve those goals. The next sections contain the definition of the optimization problem, methods for computing the objective functions, the strategy for handling constraints, the optimization algorithm and details related to the program developed.

3.1 Optimization problem definition

A normal approach in CAMD to address product design problems is optimization. In this work, the optimization problem consists of designing compounds with an optimal performance as solvent. This section describes the features of an optimal solvent for CAMD and formulates an optimization problem for solvents design.

3.1.1 Requirements for a good solvent in liquid-liquid extraction

Regardless the methodology used to identify or design the optimal solvent for a specific application, a compound must count with the next features to be considered a good candidate solvent in liquid-liquid extraction.

Extraction power K

In the process of selecting a separation solvent **s**, the target solute **t** will be recovered from the problem solvent **p**, knowing the distribution of t between s and p is of great importance. The way to quantify the extraction is with the partition ratio of t in the solvents $\mathbf{P_{t,s/p}}$ [45]. In this work this feature is just named as Extraction power **K.**

$$K = P_{t,s/p} = \frac{[t]^s}{[t]^p} \qquad \textbf{Equation 3}$$

$[t]^s$ and $[t]^p$ refer to the concentration of t in the solvent phase and problem phases, respectively. It must be mention that temperature **T** has a strong influence over the partition ratio.

In a system where the solute is in thermodynamic equilibrium with the solvents, $\mathbf{P_{t,s/p}}$ can be expressed in terms of the activity coefficients of t in s and t in p at infinite dilution, expressed as $\gamma^{\infty}_{s,p}$ and $\gamma^{\infty}_{t,m}$ respectively in the Equation 4 [46].

$$K = P_{t,s/p} = \frac{\gamma_{t,p}^{\infty}}{\gamma_{t,s}^{\infty}} \qquad \textbf{Equation 4}$$

In mass basis terms:

$$K = P_{t,s/p}\frac{MW_p}{MW_s} = \frac{\gamma_{t,p}^{\infty}}{\gamma_{t,s}^{\infty}}\frac{MW_p}{MW_s} \qquad \textbf{Equation 5}$$

In the equation, MW_p and MW_s correspond the molecular weight of the compound p and s, respectively.

Selectivity S

In a desirable scenario, the target solute moves from the problem solvent to the separation solvent without any other compound, in this case remains of problem solvent. However, in most of the cases that is not possible and another feature for evaluating separation solvents must be introduced, the Selectivity **S**. The way to estimate selectivity is by computing the solubility of t and p in the solvent s.

$$S = P_{t/p,s} = \frac{[t]^s}{[p]^s} = \frac{\gamma_{p,s}^{\infty}}{\gamma_{t,s}^{\infty}} \qquad \textbf{Equation 6}$$

In mass basis terms:

$$S = P_{t/p,s}\frac{MW_t}{MW_p} = \frac{\gamma_{p,s}^{\infty}}{\gamma_{t,s}^{\infty}}\frac{MW_t}{MW_p} \qquad \textbf{Equation 7}$$

Separation solvent loss S

In a process involving extraction operations using solvents, the recovery of the separation solvent is mandatory. Then, a third aspect to consider in the selection of a solvent is related to how much of the separation solvent gets dissolved into the problem solvent. The name of this feature is Solvent Loss **L** and it is measured by the solubility of s in p. Assuming a very low partial solubility of the binary system consisting of the problem solvent and the separation solvent, the next relationship is valid [47].

$$x_{s,p} \cdot \gamma_{s,p} = x_{s,s} \cdot \gamma_{s,s} \approx 1 \qquad \textbf{Equation 8}$$

If the concentration of the separation solvent is very close to one, the solvent loss can be expressed as follows.

$$L = x_{s,p} = \frac{1}{\gamma_{s,p}^{\infty}} \qquad \textbf{Equation 9}$$

In mass basis terms:

$$L = \frac{1}{\gamma_{s,p}^{\infty}}\frac{MW_s}{MW_p} \qquad \textbf{Equation 10}$$

3.1.2 Feasibility properties for a solvent in CAMD

At the moment of evaluating a solvent, the properties K, S and L presented above tell us how good is the solvent. However, in CAMD a candidate solvent is not selected for evaluation of the same way it is selected by a person in the traditional product design approach.

In the traditional product design approach, a solvent is selected from a list of solvents, from previous knowledge or by experience, the next step is estimating its properties and going to the laboratory in order to evaluate the performance of the solvent. On the other hand, in CAMD the solvents to evaluate are not properly "selected" based on empiric rules, these rules for selection rely on systematic and logic algorithms and can lead to compounds that are not physically feasible or that cannot be real solvents. Hence, the following properties must be taken into account in the problem of designing solvents via CAMD.

Standard Gibbs energy of formation, G_f

The Gibbs energy indicates whether a process is spontaneous or not. The Standard Gibbs energy of formation is the energy associated to the formation reaction of a compound from its consistent elements in natural state [48]. It is expressed in kJ/mol. Thermodynamics states that the more negative is the G_f for a compound, the more spontaneous its formation reaction is and the compound is more likely to exist in nature. Conversely, very positives values of G_f for a compound means that in nature the compound spontaneously would decompose into its elements. For a solvent designed by a CAMD methodology, a negative value for the standard Gibbs energy of formation ensures that the solvent can exist.

Boiling point T_b and Melting point T_m

There is no need to present a definition of these properties. This work addresses a problem of liquid-liquid separation using a solvent that must be in liquid phase at the separation conditions. The estimation of T_b and T_m let the user know the phase of the solvents generated by the CAMD methodology. The temperature of the separation T must fall in the range between T_m and T_b. That is, greater than T_m and lower than T_b.

3.1.3 Environmental concerns

A CAMD methodology is aimed to identify candidate compounds whose properties are advantageous for certain application. However, the inclusion of a candidate compound in large-scale processes can be affected if it presents any toxic behaviour for life or the environment. In addition, the world tendency is the implementation of more environment-friendly processes. To represent different effects of a compound on environment, the following properties are selected.

Fathead Minnow LC_{50}^{FM} and *Daphnia Magna* LC_{50}^{DM}

The acute toxicity indicator Lethal Concentration 50 LC_{50} stands for the concentration of a substance that is expected to be lethal to 50% of members of a tested population during a specific period of time [49]. It is expressed in mol/L. In this

work, the potential impact of a compound on the aquatic life is measured using the properties *Fathead Minnow* 96hr LC_{50}^{FM} and *Daphnia Magna* 48hr LC_{50}^{DM}.

Oral Rat LD_{50}

The acute toxicity indicator Lethal Dose 50 LD_{50} stands for the single dose of a substance that is expected to be lethal to 50% of members of a tested population from a single exposure by oral route [50]. It is expressed in mol/kg. In this work, the potential impact of a compound on the terrestrial organisms is measured using the Oral lethal dose for rats.

Water solubility W_s

This property stands for the amount of a substance that can be dissolved in liquid water. Once this amount is surpassed, the excess of substance remains in a phase apart from the water. Water solubility can be considered as an indicator in the study of environmental impact, as it tells how easy a substance can contaminate water.

Bioconcentration factor BCF

This property refers to the ratio of the concentration of a substance in aquatic biota to concentration of the substance in aqueous medium at steady-state [51]. It can be expressed in the units $\{mg/kg\}_{biota}/\{mg/l\}_{aqueous} = l_{aqueous}/kg_{biota}$, or with no units as $\{mg/kg\}_{biota}/\{mg/kg\}_{aqueous} = kg_{aqueous}/kg_{biota}$. In the same way as W_S refers to how a substance dissolves in an aquatic medium, bioconcentration factor refers to what portion of the dissolved substance is finally absorbed by aquatic organisms.

Final remarks

In the early stages of this CAMD methodology, the properties presented in the next list were taken into account, however accuracy of available estimation methods was not satisfactory [39].

- $\boldsymbol{ERA_C}$: Emission to rural air, in cases per kilogram emitted (carcinogenic).
- $\boldsymbol{ERA_{NC}}$: mission to rural air, in cases per kilogram emitted (non-carcinogenic).
- $\boldsymbol{EUA_C}$: Emission to urban air, in cases per kilogram emitted (carcinogenic).
- $\boldsymbol{EUA_{NC}}$: Emission to urban air, in cases per kilogram emitted (non-carcinogenic).

3.1.4 Market availability in CAMD

The possibility of exploring a comprehensive space for the theoretical possible compounds is one of the big benefits of using CAMD methodologies in product design problems. In this design space, optimal compounds accomplishing the physical-chemical feasibility criteria and meeting the desired features for our product can be designed.

A drawback of this approach is that not all the candidate compounds produced by CAMD are expected to be available in the market. Therefore, the designed compounds may be difficult to acquire, these have not been registered or there is no chemical pathway for the synthesis yet. In this work, the data of compounds

contained in several databases is included into the methodology in order to guide the optimization towards the design of optimal compounds available in the market. The databases used in this work are the ZINC and EPA DSSTox databases.

ZINC database

ZINC, acronym of *ZINC is not commercial*, is a public-access database originally created to be used in virtual screening, it contains millions of purchasable compounds with detailed information about the structure and the suppliers [52]. The current version of this database is ZINC15, intended to connect gene products, drugs and natural products with commercial availability. The easiest way to access the data of ZINC is via a website maintained by the University of California, San Francisco (UCSF) and the National Institute of General Medical Sciences [53].

From the list of categories present by in ZINC to classify the purchasability of compounds, Table 1 contains those used in this work. The table presents six categories, but *Wait-ok*, *Boutique* and *Annotated* are enough to classify a compound in this work.

Table 1. ZINC categories for purchasability

Category	Description	Acquisition success rate
In-stock	Ready to ship and expected delivery within 2 weeks	95%
Procurement agent	Available via procurement agents, delivery in 2 weeks	95%
Make-on-demand	Delivery typically within 8 to 10 weeks	70%
Boutique	The cost may be high but still likely cheaper than making it yourself	70%
Annotated	In catalogues but not currently for sale	--
Wait-ok	In-stock + Agent + On-demand	--

Currently ZINC counts with more than 1380 Millions of substances [54]. In this work, the extracted dataset consisted of the substances present in ZINC with a molecular weight lower than 350 Dalton and distributed as follows [55]:

- Wait-ok: 204,460,039 compound
- Boutique: 89,360,537 compounds
- Annotated: 534,234 compounds

DSSTox database

The original Distributed Structure-Searchable Toxicity (DSSTox) is a web resource maintained by the United States Environmental Protection Agency (EPA). The purpose of DSSTox intends to be an access point to high-quality data of bioassay and physicochemical data of compounds and their corresponding chemical structures [56]. DSSTox integrates the information of the following databases [57].

1. The EPA Substance Registry Services (SRS) database
2. The National Library of Medicine's (NLM) ChemID-Plus

3. Part of the PubChem database of the National Center for Biotechnology Information (NCBI), corresponding to approximately a dataset of 700,000 entries.

The last version of the DSSTox database can be accessed via the CompTox Chemistry Dashboard website [58]. CompTox is a curated and open-access resource containing more than 720,000 chemicals of relevance for environmental studies [59]. The number of compounds extracted for this work is of 720,839.

Market availability category, A

Based on the databases introduced above, the CAMD methodology of this work classifies the market availability of a compound as follows:

1. *Available*: The compound is present either in the DSSTox database or in the Wait-ok subset of ZINC.
2. *Hard-to-acquire*: The compound is present either in the subset Boutique or Annotated of ZINC.
3. *Unavailable*: The compound is not present in any of the databases.

The next section describes the strategy used by the CAMD methodology to include these categories into the solvent design process.

3.1.5 Designing the best separation solvent

To wrap it all up, a formal definition of the multi-objective optimization problem of the CAMD methodology proposed in this work is presented below.

3.1.5.1 Objective functions

Compared with the original CAMD methodology proposed in by Serrato, one of the major improvements of this work is the multi-objective approach. However, the original optimization function of that work [4] is still present in the methodology of this work as one of the objectives functions.

1. **Maximize the Solvent yield Y**

$$\max \mathrm{Y} \qquad \textbf{Equation } 11$$

$$\mathrm{Y} = \mathrm{KS} = \frac{\gamma_{\mathrm{t,p}}^{\infty}}{\gamma_{\mathrm{t,s}}^{\infty}} \frac{\gamma_{\mathrm{p,s}}^{\infty}}{\gamma_{\mathrm{t,s}}^{\infty}} \frac{\mathrm{MW_t}}{\mathrm{MW_s}} \qquad \textbf{Equation } 12$$

2. **Minimize the Environmental index E**

$$\min \mathrm{E} \qquad \textbf{Equation } 13$$

$$\mathrm{E} = \mathrm{k_{WS}} \log \mathrm{W_s} + \mathrm{k_{BCF}} \log \mathrm{BFC} - \mathrm{k_{LD50}} \log \mathrm{LD_{50}} - \log \mathrm{LC_{50}^{FM}} - \log \mathrm{LC_{50}^{DM}} \qquad \textbf{Equation } 14$$

Regarding Equation 14, the method used to obtain the coefficients k_{WS}, k_{BCF} and k_{LD50} relies on data obtained from the computation of *Fathead Minnow* $\mathrm{LC_{50}}$, *Daphnia Magna* $\mathrm{LC_{50}}$, Oral Rat $\mathrm{LD_{50}}$, Bioconcentration factor and Water solubility

for 50 000 compounds randomly created. Standardization factors for each property are calculated and transformed in such a way that all properties vary within the same scale as *Fathead Minnow* LC_{50} and *Daphnia Magna* LC_{50} do. Section 4.1.3 explains in detail the considerations taken into account in this process, and how Equation 14 is obtained.

3.1.5.2 Constraints

Most of the constraints related with thermo-physical properties are those defined in the work of Serrato [4].

- **Melting point**
  During an extraction process, solvent and solute must be in liquid phase, then the melting point for the designed solvents must be below the operation conditions. The temperature in standard conditions is 25ºC (298K) and this work uses a slightly lower value of temperature of 20ºC (293K) as the upper limit for melting point.

$$T_m < 293K \qquad \textbf{Equation } \textit{15}$$

- **Boiling point**
  As in the case of melting point, the boiling point is intended to maintain the solvents in liquid phase. In addition, the boiling point must ease the later process of separation where the solute to be extracted is separated from the separation solvent and the separation solvent is recovered to be reused in a new separation. To cover both situations, the designed solvents in this work have a boiling point lower than 300ºC (573K).

$$T_b < 573K \qquad \textbf{Equation } \textit{16}$$

- **Standard Gibbs energy of formation**
  As previously mentioned, a compound with positive value for the Gibbs energy of formation is not thermodynamically feasible. That statement could lead to propose a positive value of this property as one constraint in the CAMD methodology. However, in nature, some compounds do not decompose despite of having a small positive value of Gibbs energy of formation. To give these compounds a chance to be selected as candidates, the constraints for this property allow small positive values.

$$\Delta G_f < 100\ kJ / mol \qquad \textbf{Equation } \textit{17}$$

- **Solvent loss**[2]
  The maximum permissible solvent loss in this work is 10%. It means that the maximum acceptable concentration of the separation solvent in the problem solvent is of 0.1 mole fraction.

$$L < 0.1 \qquad \textbf{Equation } \textit{18}$$

- **Market availability**

[2] In Section 4.3 of results, the impact of this constraint on the solvents designed by the CAMD methodology is discussed and the constraint is redefined.

According to the market availability categories introduced in section 3.1.4, a solvent *s* meets the first constraint if it is present in the *available* or *hard-to-acquire* category.

$$A = s \in \{C_{available} \cup C_{hard-to-acquire}\} \quad \textbf{Equation } \textit{19}$$

3.2 THERMO-PHYSICAL PROPERTIES

The estimation of the thermo-physical properties introduced in section 3.1.2 and others included in this work is done using the group-contribution method proposed by Hukkerikar [60]. In this model, properties can be predicted based on the first-order, second-order and third-order group contributions present in a molecule. In this work, only first-order and second-order contributions are considered. Third-order contributions are excluded as these contributions are intended to allow the estimation of complex heterocyclic and large poly-functional acyclic chemicals (C=7 to 60) which are not covered in this CAMD methodology. In line with the mentioned considerations, the next formula is used for predicting thermo-physical properties.

$$f(X) = \sum_i N_i C_i + \sum_j N_j C_j \quad \textbf{Equation } \textit{20}$$

In the equation, f(X) is a function for the property X to be predicted. N_i and N_j are the number of occurrences for the first-order and the second-order groups in a molecule, respectively. C_i and C_j are the respective contributions of the first-order and second-order groups to a property. Values for C_i and C_j are the reported in the step-wise parameter tables for thermo-physical properties in the work of Hukkerikar [61].

Table 2. GC equations for thermo-physical properties prediction [61]

Property	Equation, f(X)	Constants	Units
Standard Gibbs energy of formation, G_f	$\exp\left(\frac{T_b}{T_{b0}}\right)$	$T_{b0} = 244.5165K$	[K]
Normal boiling point T_b	$\exp\left(\frac{T_m}{T_{m0}}\right)$	$T_{m0} = 143.5706K$	[K]
Normal melting point T_f	$G_f - G_{f0}$	$G_{f0} = -1.3385 \frac{kJ}{mol}$	$\left[\frac{kJ}{mol}\right]$
Liquid molar volume V_m at 298K	$V_m - V_{m0}$	$V_{m0} = 0.0160 \frac{cc}{kmol}$	$\left[\frac{cc}{kmol}\right]$

According to Equation *20*, once all the group-contributions for a molecule are found to certain property X, the function f(X) allows to compute the value of the property. Table 2 summarizes the functions f(X) of the thermo-physical properties predicted in this work. As a complement to this table, it is worthwhile to mention that the density ρ of a compound can be computed using the molecular weight of and the liquid molar volume as follows:

$$\rho = \frac{MW}{V_m}$$ **Equation *21***

3.3 ENVIRONMENT-RELATED PROPERTIES

The estimation of the environment-related properties introduced in section 3.1.3 and included in this work is done by using the group-contribution method proposed by Hukkerikar [39]. The approach of this model is the same used for predicting of thermo-physical properties presented above and the same considerations are applied in this case. That is to say, only first-order and second-order contributions are considered and Equation *20* can also be used to predict environment-related properties. Values for C_i and C_j are the reported in the step-wise parameter tables for environment-related properties in the work of Hukkerikar [61].Table 3 summarizes the functions f(X) of the environment-related properties to be predicted in this work.

Table 3. GC equations for environment-related properties prediction [61]

Property	Equation, $f(X)$	Constants	Units
Fathead Minnow LC50[3], LC_{50}^{FM}	$-\log LC_{50}^{FM} - FM_0$	$FM_0 = 2.1949$	$\left[\frac{mol}{lit}\right]$
Daphnia Magna LC50[4], LC_{50}^{DM}	$-\log LC_{50}^{DM} - DM_0$	$DM_0 = 2.9717$	$\left[\frac{mol}{lit}\right]$
Oral Rat LD_{50}	$-\log LD_{50} - A_{LD50} - B_{LD50}MW$	$A_{LD50} = 1.9372$ $B_{LD50} = 0.0016$	$\left[\frac{mol}{kg}\right]$
Water solubility, W_s	$\log W_s - A_{WS} - B_{WS}MW$	$A_{WS} = 4.5484$ $B_{WS} = 0.3411$	$\left[\frac{mg}{lit}\right]$
Bioconcentration factor, BCF	$\log BCF$	-	-

3.4 MARKET AVAILABILITY

Section 3.1.4 listed the sources used in the CAMD methodology in this work to assign market availability of compounds. The process of extraction, pre-processing and insertion of the compounds data into a single database is described below.

[3] In the original source, this equation is introduced as $-\log LC_{50}^{FM} + FM_0$, and some works uses it [95]. Nevertheless, in this work the equation was slightly changed to produce results more approximate to the experimental validation data.

[4] In the original source, this equation is introduced as $-\log LC_{50}^{DM} + DM_0$. Nevertheless, in this work the equation was slightly changed to produce results more approximate to the experimental validation data.

3.4.1 Database model

A MongoDB instance is selected as the database engine to store the compounds queried by the CAMD methodology. MongoDB is an open-source document database capable to store unstructured data (NoSQL) as *documents* inside *collections* [62]. In this CAMD methodology, a *Compounds* collection is created to store the compounds that are used to assign the market availability of the designed solvents. For a compound document, the minimal fields in the database collection are the following:

- Source: the source of origin defines whether the compound is *Available*, *Hard-to-acquire* or *Unavailable* as defined in 3.1.4.
- SMILES: the chemical structure according to the SMILES notation. This is the most important field as each compound generated in the optimization process will use this field to query the database.
- Id: the identifier of the compound in the original source (ZINC or DSSTox), this field is used to link the compound with a website containing the available vendors.

Simplified Molecular Input Line System, SMILES

SMILES is a chemical notation language designed to represent molecular structures as a linear string of symbols, similar to natural language [63]. SMILES is intended to be easily interpreted by chemists and computer systems without ambiguities. The advantages of using SMILES in molecule representation are the following:

1. The line notation used allows the description of unique chemical graphs comprising nodes (atoms) and edges (bonds).
2. The structure specification is user-friendly and the input rules can be learned quickly and naturally.
3. The interpretation is a machine-friendly and machine-independent system capable of generating or interpreting unique structures of any complexity.

In the SMILES notation there is no spaces. Hydrogen atoms may be omitted or included. SMILES encoding rules comprises specifications for the representation of atoms, bonds, branches, ring closures, disconnections, isomeric forms and chiral centres [64].

The flexibility of SMILES notation makes possible for a molecule to hold a large number of valid representations. A simple linear molecule such as ethanol (CH_3-CH_2-OH) can be parsed either as CCO or as OCC (and of many other forms), the name of this encoding is Generic SMILES. In contrast, the Canonical SMILES of a molecule is a unique SMILES among all the valid possibilities. In software tools, the implementation of methods to parse molecules and map SMILES must be based on canonicalization algorithms. Currently there is no universally accepted method to canonize SMILES, a variety of canonicalization algorithms are implemented in software tools and open source projects [65].

3.4.2 ZINC database

Information about the compounds registered in ZINC can be queried and downloaded using *tranches*. In ZINC, a tranche is a text file containing the locations (as URLs) of subsets of compounds data ready to download. The ZINC website counts with a *Tranche Browser* section where the tranches can be filtered according to different criteria [66]. For this work, the filters used to build the market availability database are *LogP* (octanol–water partition coefficient), *Purchasability* and *Molecular Weight*.

To handle market availability in this CAMD methodology, data of compounds with any value of LogP and a molecular weight lower than 350 Da (equivalent to a C14-alkane chain) are downloaded for the purchasability criteria *Wait-Ok*, *Annotated* and *Boutique*. Figure 2, Figure 3 and Figure 4 display the number of compounds in the tranches processed for this work, these correspond respectively to the purchasability criteria *Wait-Ok*, *Annotated* and *Boutique*.

	Molecular Weight (up to, Daltons)					
LogP (up to)	200	250	300	325	350	3
-1	18,093	90,865	275,487	404,529	778,332	
0	132,954	861,904	2,477,076	2,555,669	4,607,804	
1	409,556	3,415,953	11,228,683	10,649,404	17,840,904	
2	566,541	6,138,302	25,673,560	25,756,846	22,959,662	
2.5	211,290	3,012,557	15,741,531	16,961,722	17,520,288	
3	115,889	2,253,291	14,298,657	16,590,843	27,077,211	
3.5	45,494	1,288,623	10,272,185	13,166,174	21,890,997	
4	9,195	420,318	5,301,653	7,885,082	12,418,918	
4.5	2,266	54,938	1,627,929	3,349,787	6,264,859	
5	285	4,348	240,431	813,460	2,036,809	
>5	30	2,532	22,971	111,136	407,301	
Totals, by Weight	1,511,593	17,543,631	87,160,163	98,244,652	0	

Figure 2. Wait-Ok tranches sizes.

	Molecular Weight (up to, Daltons)					
LogP (up to)	200	250	300	325	350	3
-1	150,063	356,286	342,674	47,063	9,248	
0	588,523	1,585,982	1,630,799	211,711	36,865	
1	1,556,873	4,734,301	5,239,675	678,807	198,319	
2	2,129,206	8,308,009	10,782,097	1,558,107	779,049	
2.5	841,538	4,410,424	7,033,840	1,169,731	773,197	
3	492,971	3,677,132	6,924,618	1,372,265	1,048,854	
3.5	221,042	2,432,765	5,949,139	1,404,678	1,262,753	
4	70,905	1,280,951	4,409,006	1,259,132	1,239,591	
4.5	9,920	512,181	2,468,655	966,950	1,040,673	
5	1,357	131,272	1,113,423	512,659	664,465	
>5	316	18,352	459,271	315,868	408,183	
Totals, by Weight	6,062,714	27,447,655	46,353,197	9,496,971	0	

Figure 3. Boutique tranches sizes.

	Molecular Weight (up to, Daltons)					
LogP (up to)	200	250	300	325	350	3
-1	1,312	3,323	7,234	3,519	5,230	
0	2,017	4,156	7,703	3,485	4,058	
1	4,247	9,661	17,529	8,606	7,859	
2	5,169	17,929	42,418	21,680	24,492	
2.5	1,790	11,293	31,870	19,686	22,874	
3	1,130	11,004	36,510	26,799	30,286	
3.5	646	8,537	36,078	32,660	38,507	
4	208	6,013	28,473	31,469	42,259	
4.5	102	2,813	17,605	26,514	37,275	
5	43	1,245	8,537	15,980	26,601	
>5	17	331	5,821	11,070	23,506	
Totals, by Weight	16,681	76,305	239,778	201,468	0	

Figure 4. Annotated tranches sizes.

Importing compounds into the market availability database

The structure of each compound of the ZINC database is encoded as SMILES according to the Canonical form implemented in the Open source toolkit for cheminformatics RDKit [67]. Once the tranches subsets are downloaded, the pre-processing step for the compounds consists in reparsing the original ZINC Canonical SMILES to the Canonical SMILES implemented in The Chemistry Development Kit 2.0 (CDK) [68], the parameter *SmiFlavor* flag in the parsing method is set to *Absolute*, this parameter defines the strategy to be used for handling stereochemistry issues. Compounds with reparsed SMILES are inserted into the *Compounds* collection of the MongoDB database. Figure 5 illustrates the process of importing compounds from ZINC to the Market Availability database.

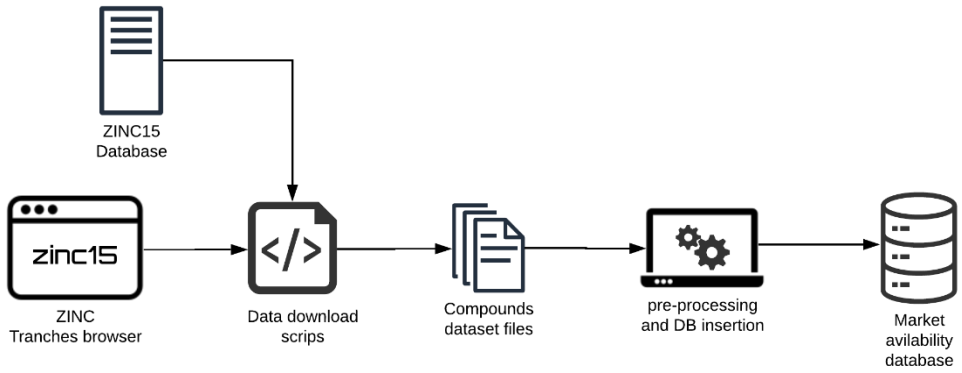


Figure 5. ZINC data import process.

3.4.3 DSSTox database

The acquisition of the DSSTox dataset is easier compared with the ZINC datasets. While the former has just more than is 700 thousand compounds, the latter has around 300 million. As mentioned in section 3.1.4, the CompTox Chemistry Dashboard website contains the DSSTox dataset. The version used in this work is the *DSSTox MS Ready Mapping File* of April 2018. The DSSTox dataset consists of several MS Excel files with the information of the compounds, each workbook contained 200,000 compounds.

Importing compounds into the market availability database

The structure of each compound of the DSSTox database is encoded as SMILES and as InChI. The latter corresponds to the IUPAC International Chemical Identifier [69]. Once the MS Excel workbooks with the subsets are downloaded, the pre-processing step consists in converting the workbooks files into Comma-Separate Values format (csv) and reparsing the SMILES of each compound to the Canonical SMILES implemented in The Chemistry Development Kit 2.0 (CDK) [68], the *SmiFlavor* chosen for this process is *Absolute*. The compounds with the new SMILES are inserted into the *Compounds* collection of the MongoDB database. Figure 6 illustrates the process of importing compounds from the CompTox website to the Market Availability database.

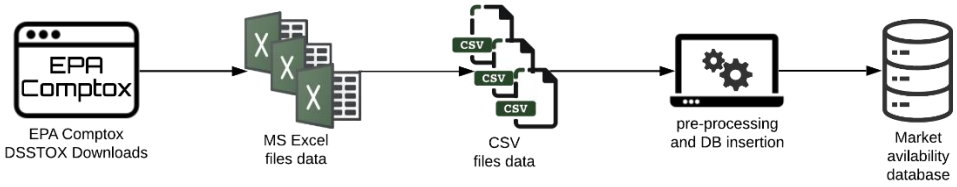


Figure 6. DSSTox data import process.

3.5 MIXTURE PROPERTIES

The mixture properties used in this CAMD methodology correspond to the activity coefficients required for the estimation of the properties presented in section 3.1.1. The method used for the estimation of activity coefficients is the UNIFAC-Dortmund (Do) method, which is an extension of the original UNIFAC. In the development of this method, data from the Dortmund Data Bank, the worldwide largest factual databank for thermo-physical properties, are used in the regression of the interaction parameters [70]. However, the main difference between the UNIFAC-Do and the original UNIFAC method is that the former introduces a more rigorous equation for the estimation of the temperature-dependent group interaction parameters. In addition, as the UNIFAC method presents a poor performance for the prediction of activity coefficients at infinite dilution (γ^{∞}), the UNIFAC- Do method took into account the activity coefficients at infinite dilution of many more mixtures in the regression process of the group-group interaction parameters [71].

In UNIFAC, as in all the group-contribution methods, a mixture of molecules is represented as a mixture of pre-defined structural units, the so-called *solution-of-groups* concept. Each structural unit is named *subgroup* and makes part of a *functional group* [72] and the subgroups in a functional group keep some relationship (e.g. these belong to alkenes, carboxyl, ...). Computing activity coefficients using UNIFAC requires the identification of the functional groups and subgroups comprising the mixture. Figure 7 illustrates an example of this concept.

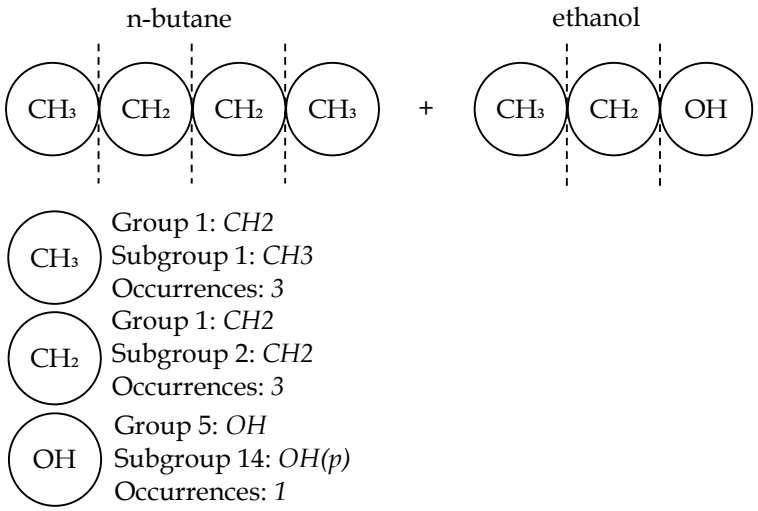


Figure 7. Representation of a mixture as group-contributions.

The estimation of the activity coefficient for the compound i, which is part of a mixture of NC components, is computed as the sum of a combinatorial (C) and a residual (R) part [73].

$$\ln \gamma_i = \ln \gamma_i^C + \ln \gamma_i^R \qquad \textbf{Equation 22}$$

The combinatorial part accounts for the differences in the size and shape of the molecules, while the residual accounts for the contribution of intermolecular forces derived from the solution-of-groups to the non-ideality of the mixture.

The combinatorial part for the compound i is estimated as follows.

$$\ln \gamma_i^C = 1 - V_i' + \ln V' - 5q_i\left(1 - \frac{V_i}{F_i} + \ln \frac{V_i}{F_i}\right) \qquad \textbf{Equation 23}$$

Where V_i stands for the volume fraction ratio of the component i, F_i for the surface fraction ratio of the component i and V'_i is an improvement of the volume fraction ratio introduced by the UNIFAC-Dortmund method.

$$V_i = \frac{r_i}{\sum_j^{NC} r_j\, x_j} \qquad \textbf{Equation 24} \qquad F_i = \frac{q_i}{\sum_j^{NC} q_j\, x_j} \qquad \textbf{Equation 25}$$

$$V_i' = \frac{r_i^{3/4}}{\sum_j^{NC} r_j^{3/4}\, x_j} \qquad \textbf{Equation 26}$$

The molecular van der Waals volume r_i and molecular surface area q_i of compound i present in the equations above are calculated using the next equations.

$$r_i = \sum_k v_k^i\, R_k \qquad \textbf{Equation 27} \qquad q_i = \sum_k v_k^i\, Q_k \qquad \textbf{Equation 28}$$

In the equations, v_k^i is the number of subgroups of type k in the compound i, while R_k and Q_k are the van der Waals volume and surface area of the subgroup k. In this work, the values used for R_k and Q_k are taken from the Dortmund Data Bank, maintained and updated by the DDBST GmbH group [74].

On the other hand, the residual part for the compound i is estimated as follows.

$$\ln \gamma_i^R = \sum_k v_k^i(\ln \Gamma_k - \ln \Gamma_k^i) \qquad \textbf{Equation 29}$$

Where Γ_k stands for the residual activity of group k in a mixture with all the NG groups, and Γ_k^i stands for the residual activity of group k in a mixture containing only compound i. Both terms are calculated using the next equations.

$$\ln \Gamma_k = Q_k\left[1 - \ln\left(\sum_n^{NG} \theta_n \Psi_{nk}\right) - \sum_n^{NG} \frac{\theta_l \Psi_{kn}}{\sum_m^{NG} \theta_m \Psi_{mn}}\right] \qquad \textbf{Equation 30}$$

$$\ln \Gamma_k^i = Q_k\left[1 - \ln\left(\sum_n^{NG} \theta_n^i \Psi_{nk}\right) - \sum_n^{NG} \frac{\theta_n^i \Psi_{kn}}{\sum_m^{NG} \theta_m^i \Psi_{mn}}\right] \qquad \textbf{Equation 31}$$

Where θ_n stands for the area fraction ratio of the subgroup n respect to all the subgroups in the mixture, while θ_n^i for the area fraction ratio of the subgroup n respect to the subgroups of compound i.

$$\theta_n = \frac{Q_n X_n}{\sum_m^{NG} Q_m X_m} \quad \textbf{Equation 32} \qquad \theta_n^i = \frac{Q_l X_n^i}{\sum_m^{NG} Q_m X_m^i} \quad \textbf{Equation 33}$$

$$X_m = \frac{\sum_i^{NC} v_m^i x_i}{\sum_j^{NC} \sum_n^{NG} v_n^j x_j} \quad \textbf{Equation 34} \qquad X_m^i = \frac{Q_l X_l}{\sum_j^{NC} \sum_n^{NG} v_n^j x_j} \quad \textbf{Equation 35}$$

In the equations above, X_m is the fraction ratio of the subgroups of type m in the mixture, while X_m^i is the fraction ratio of the subgroups m in the compound i. In the case of Ψ_{nm}, this term stands for group-group interactions. It is temperature dependent and is calculated using the next equation introduced in UNIFAC Dortmund.

$$\Psi_{nm} = \exp\left(-\frac{a_{nm} + b_{nm}T + c_{nm}T^2}{T}\right) \quad \textbf{Equation 36}$$

In the application of this equation, user must keep in mind that the n and the m are not interchangeable, i.e. $a_{nm} \neq a_{mm}$, $b_{nm} \neq b_{mn}$ and $c_{nm} \neq c_{mn}$.Parameters a_{nm}, b_{nm} and c_{nm} can be found in tables. In this work, these parameters as well as R_k and Q_k were taken from the Dortmund Data Bank [74].

3.6 Optimization strategy

The solution of a CAMD problem can be challenging, as its mathematical representation can lead to dimensionally high or strongly non-linear models. Nevertheless, meta-heuristics approaches for optimization have proven to be useful in CAMD as these do not require derivative information, the implementation is relatively easy and these can be adapted to a wide type of problems [75]. In addition, single-objective and multi-objective problems can be addressed with these methods. As a consequence, meta-heuristic approach has become popular in CAMD methodologies today.

The previous sections introduced all the required methods to understand the solvent design problem and the evaluation of the performance of a solvent. This section explains how the optimization stage is conducted.

3.6.1 Selected optimization algorithm

In line with the comments above, a multi-objective metaheuristic evolutionary algorithm has been chosen to solve the optimization problem contained in this CAMD methodology. The selected algorithm is a multi-objective adaptation of the metaheuristic evolutionary algorithm HAEA [76]. Below is presented a description of the original single-objective HAEA and the multi-objective version of that algorithm.

3.6.1.1 Hybrid Adaptive Evolutionary Algorithm (HAEA)

HAEA is a genetic algorithm whose conception differs from traditional genetic algorithms. In traditional algorithms, the appearance of new offsprings depends of genetic operators with static rates of application over the population. Besides of the little flexibility, the solution of a problem using the traditional approach requires the time-consuming task of tuning genetic operator rates. Alternatively, evolution of individuals in HAEA is guided by variable genetic operator rates that evolve independently for each individual along the progress of the algorithm.

In HAEA, the introduction of genetic operator rates by the user is not necessary as the algorithm set the operator rates for new offsprings according to the performance of the operators applied to the parents. In the initialization of the algorithm, the rate for every operator is set as one divided the number of operators and the rates of an individual are inherited to its offspring. An individual cannot produce an offspring of a size larger than one and in the case of operators with offspring larger than one, a tournament is performed to choose the best child. Finally, HAEA allows the easy inclusion of custom genetic operators adapted to a particular problem. Algorithm 1 is a summary of HAEA.

Algorithm 1. Hybrid Adaptive Evolutionary Algorithm (HAEA) [76]

```
HAEA ( λ, termination_condition )    input: population size λ, termination condition
 1. t_0 = 0
 2. P_0 = initPopulation( λ ) ,
 3. while(termination_condition( t , P_t ) is false ) do
 4.   P_t+1 = {}
 5.   for each ind ∈ P_t do
 6.     rates = operators_rates[ind]
 7.     δ = random(0,1)     // learning rate
 8.     oper = SELECT_OPERATOR( operators, rates )
 9.     parents = SELECT_PARENTS( P_t , ind )
10.     offspring = apply( oper, parents )
11.     child = BEST ( offspring , ind )   // Best child according to the fitness
12.       if( FITNESS( child ) > FITNESS( ind ) ) then
13.         rates[oper] = (1.0 + δ) * rates[oper]   // reward
14.       else
15.         rates[oper] = (1.0 - δ) * rates[oper]   // punish
16.     normalize rates( rates )
17.     operators_rates[child] = rates
18.   P_t+1 = P_t+1 ∪ {child}
19. t = t + 1
```

3.6.1.2 Multi-Objective Hybrid Adaptive Evolutionary Algorithm, MOHAEA

In the development of single-objective optimization algorithms, most of the attention is focused on convergence. The one-dimensional space where individual fitnesses advance makes easy to estimate this metric. In this cases, the relationship between the fitness and the objective function is very tight. In a single-objective optimization problem, the fitness is usually the same objective function, unless the problem is constrained.

On the other hand, the convergence in multi-objective algorithms must be measured in a n dimensional space, according to the n objective functions. In these cases, the introduction of Pareto dominance and the Pareto optimal front may be necessary if an unbiased solution is desired. One of the major challenges in the development of multi-objective optimization algorithms is to ensure not only convergence but diversity in the solutions. In the context of multi-objective optimization, both are defined as follows [77]:

- Convergence: or accuracy, refers to how distant are the optimal solutions found from the theoretical (or known) Pareto optimal front.
- Diversity: it refers to how is the distribution and the spread of the solutions. Distribution is the relative distance between solutions, while the spread is the range of values covered by the solutions.

The MOHAEA implementation proposed in this work addresses convergence and diversity using an approach based on the fast non-dominated sorting and the crowding-distance introduced for first time in the NSGA-II algorithm [78]. The algorithm presented below was developed in collaboration with Juan Camilo Castro Pinto, M.Sc. Student of Computer and Systems Engineering of the Universidad Nacional de Colombia.

Fitness assignment

In MOHAEA, the performance of individuals is measured using only fitness. For the individuals of a population, fitness assignment requires sorting the population according to non-domination levels and computing the crowding distance of the individuals of each level. The fitness of an individual corresponds to the sum of the non-domination rank and the crowding-distance as shown below.

$$\mathrm{FITNESS(i)} = \mathrm{i}_{\mathrm{rank}} + \mathrm{i}_{\mathrm{distance}} \qquad \textbf{Equation 37}$$

Where i is the individual, $\mathrm{i}_{\mathrm{rank}}$ and $\mathrm{i}_{\mathrm{distance}}$ are respectively the ranking and the crowding distance of the individual. The benefits of this approach is that produces a single measure covering convergence and diversity despite of the number of objective functions in the problem.

Ranking assignment

The rank of an individual is defined according to the Fast Non-Dominated Sorting approach presented in the NSGA-II algorithm [78]. The detailed procedure is presented in Algorithm 2. The purpose of Fast Non-Dominated Sorting is to extract the different levels of non-dominance of a population. The Pareto dominance operator ($\prec$) is used to separate each front. The first level consists of the non-dominated individuals, the second level comprises the individuals dominated only by lower level individuals, and so on. The output of Algorithm 2 is a list of the non-domination ranks $\mathcal{F}$, each rank is a list of the individuals classified on the respective Pareto front.

Algorithm 2. Fast Non-Dominated Sorting [78]

```
NON_DOMINATED_SORT( P )    input: population P
 1. F = {}
 2. for each p in P do
 3.   S_p = {}   // set of solutions dominated by p
 4.   n_p = 0   // domination count, number of solutions that dominates p
 5.   for each q in P do
 6.     if p ≺ q then    // if p dominates q
 7.       S_p = S_p{q}    // q is added to the solutions dominated by p
 8.     else if q ≺ p then
 9.       n_p = n_p + 1   // increment in the domination counter of p
10.   if n_p = 0 then
11.     p_rank = 1
12.     F_1 = F_1 ∪ {p}   // p belongs to the Pareto front 1
13. i = 1 // initialization of front counter
14. while F_i ≠ {}
15.   Q = {}   // individuals of next front
16.   for each p in F_i
17.     for each q in S_p
18.       n_q = n_q − 1
19.       if n_q = 0 then   // if all the members dominating q have been removed
20.         q_rank = i + 1
21.         Q = Q ∪ {q}
22.   i = i + 1
23.   F_i = Q
24. return F
```

Distance assignment

The distance contribution is assigned based on the *crowding-distance* diversity measure presented in the NSGA-II algorithm [78]. This indicator is an estimation of the density of solutions surrounding a particular solution in a Pareto front.

The crowding-distance for a solution is calculated as the Manhattan distance formed by the solution and the nearest surrounding neighbours of the same Pareto front, Figure 8 shows how are selected the neighbours of the solution *i*, in the case that there are no neighbours surrounding the solution, the distance takes the maximum value of 1.0 for each dimension of the objectives space. Algorithm 3 presents the process to compute the crowding-distance of the individuals of a front.

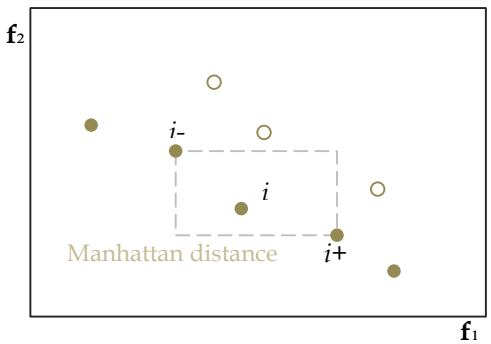


Figure 8. Crowding-distance calculation

Algorithm 3. Crowding-distance assignment for a Pareto front

```
ASSIGN_CROWDING_DISTANCES( ℐ )    input: Pareto front ℐ
1. l_ℐ = |ℐ|   // number of individuals in ℐ
2. for each i in ℐ do
3.   ℐ[i]_crowd−dist = 0   // distance initialization
4. for each m in problem_objectives do
5.   SORT( ℐ, m )   // sort ℐ according to objective m values
6.   ℐ[1]_crowd−dist = ℐ[1]_crowd−dist + 1   // maximum distance for the boundary points
7.   ℐ[l_ℐ]_crowd−dist = ℐ[l_ℐ]_crowd−dist + 1
8.   for i = 2 to ( l - 1 ) do
9.     ℐ[i]_crowd−dist = ℐ[i]_crowd−dist + (ℐ[i + 1]. m − ℐ[i − 1]. m) / (ℐ[l_ℐ]. m − ℐ[1]. m)
```

Once the crowding-distance $i_{crowd-dist}$ of an individual has been computed, the distance contribution to the fitness is calculated as shown in Equation *38*. It must be noted that the value of $i_{distance}$ is always positive and lower than one. This feature makes possible for individuals of the same Pareto rank to have fitness values (Equation *37*) within a well-defined range that does not overlap the ranges of the other Pareto ranks.

$$i_{distance} = \frac{1}{2 + i_{crowd-dist}} \qquad \textbf{Equation } \textit{38}$$

Evolution

The main loop of MOHAEA starts with the creation of an initial population P_0 of size λ and the assignment of genetic operator rates evenly for each individual of the population. As well as HAEA does, an individual is evolved using a genetic operator selected by roulette-wheel according to the operator rates (the *SELECT_OPERATOR* method). The selected operator is applied to the individual and if other individuals are required in this process, these are selected from the population using any selection strategy (*SELECT_PARENTS* method). Once the offspring of a parent individual has been produced, the children are included in the *offsprings* population.

The process of selecting the operator for an individual and producing an offspring must be repeated for all the individuals in the population. It must be pointed out that according to the genetic operator used, an individual can produce more than one children, resulting in a variable size for the offsprings population. Nevertheless, the minimum number of individuals in this population is the size λ of the current population.

The selection of the individuals of the next generation starts with the combination of the current population and the offsprings population into a *combined* population. Afterwards, fitness values are assigned for all the individuals in the combined population.

The fitness of an individual is computed using Equation *37* as shown above. The application of this equation requires to know the Pareto rank and the crowding-distance of each individual. The values for both indicators are assigned using the

Fast Non-Dominated Sorting and Crowding-Distance procedures presented in Algorithm 2 and Algorithm 3, respectively.

The next step is the selection of the λ individuals from the combined population that will be part of the next population. The *BEST_BY_FAMILY* operator is the method used to select these individuals. This operator iterates over the "families" present in the combined population and the selection of the best individual of each one according to the assigned fitness. A family consists of the parent that produced some children, the parent is an individual of the current generation and the children are the offspring. The individual with the best fitness in a family is included in the *next* population.

Algorithm 4. Selection of the best individuals

```
BEST_BY_FAMILY ( families , operator_rates ) input: families, map of genetic operator rates
 1. new_population = {}
 2. for each f in families do
 3.   SORT( f_children , 'fitness' )   // sorting offspring according to the fitness
 4.   best_child = f_children[1]   // child with the best fitness
 5.   go_rates = operators_rates[f_parent]   // genetic operator rates of the parent
 6.   oper = f_operator   // genetic operator used to generate the offspring
 7.   δ = random(0,1)   // learning rate
 8.   best_individual = {}
 9.   if FITNESS( best_child ) < FITNESS( f_parent ) then   // is the best child better?
10.     go_rates[oper] = (1.0 + δ) * go_rates[oper]   // reward
11.     best_individual = best_child
12.   else
13.     go_rates[oper] = (1.0 - δ) * go_rates[oper]   // punish
14.     best_individual = f_parent
15.   go_rates = NORMALIZE( go_rates )
16.   operators_rates[ best_individual ] = ind_rates
17.   new_population = new_population ∪ { best_individual }
18. return new_population
```

The *BEST_BY_FAMILY* operator also updates the genetic operator rates of the same way HAEA does. A parameter δ with a random value between 0 and 1 is created and depending on whether the best individual of the family is the parent or one of the children the δ values is added to or subtracted from the rate of the genetic operator used to generate the child. If the best individual is a child δ is added and, conversely, if the best individual is the parent, δ is subtracted. The updated genetic operator rates are normalized and assigned to the best individual of the family. Algorithm 4 shows the process to apply this operator for a set of families.

Once the next population has been filled out and the genetic operator rates have been updated, the new population replaces current generation and a new iteration of the algorithm is performed again less the *TERMINATION_CONDITION* had been met. Algorithm 5 collects the methods described above and summarizes the proposed MOHAEA algorithm.

Algorithm 5. Multi-Objective Hybrid Adaptive Evolutionary Algorithm (MOHAEA)

```
MOHAEA ( λ, termination_condition )
 1. t_0 = 0
 2. P_0 = INIT_POPULATION( λ )
 3. while TERMINATION_CONDITION( t , P_t ) is false do
 4.   families = {}
 5.   offsprings = {}
 6.   for each ind in P_t do
 7.     ind_rates = operators_rates[ind]
 8.     oper = SELECT_OPERATOR( operators, ind_rates )
 9.     parents = SELECT_PARENTS( P_t )
10.     ind_offspring = APPLY_OPERATOR( oper , parents )
11.     offsprings = offsprings ∪ { ind_offspring }
12.     f = {}   // familiy
13.     f_children = ind_offspring
14.     f_parent = ind
15.     f_operator = oper
16.     families = families ∪ { f }
17.   combined_pop = P_t ∪ offsprings
18.   F_t = NON_DOMINATED_SORT( combined_pop )
19.   for each front in F_t do
20.     ASSIGN_CROWDING_ DISTANCES( front )   // to compute fitness
21.   P_t+1 = BEST_BY_FAMILY ( families , operator_rates )
22.   t = t + 1
```

Constrains handling

In the process of designing optimal compounds using the CAMD methodology of this work, the most important step is the solution of the constrained multi-objective optimization problem introduced in Section 3.1.5, hence MOHAEA needs a strategy to handle constrained problems. The approach selected is based on the constraint-domination approach introduced by the NSGA-II algorithm [78]. Under this approach any feasible individual has a better non-domination rank compared with any unfeasible individual. This approach is powerful and easy to apply. However, a drawback of constraint-domination is that it makes difficult for an individual located near to the feasibility region to participate in the generation of new offsprings [79].

MOHAEA uses a variation of the constraint-domination approach in order to make it more flexible and provide unfeasible solutions more opportunities to produce feasible solutions. The new approach consists in that any unfeasible individual has a worse non-domination rank compared with the feasible individuals of the same Pareto rank. In terms of performance, an individual is reduced by one Pareto front for each violated constraint. As the next equation shows, this is achieved by adding one unit to the fitness value for every constraint violated.

$$\mathrm{FITNESS}(i) = i_{rank} + i_{distance} + \sum_{c=1}^{C} W_c(i) \qquad \textbf{Equation } 39$$

Where C is the number of constraints and $W_c(i)$ is the weight of the constraint c evaluated on the individual i defined as follows.

$$Wc(i) = \begin{cases} 0 & \text{if i does not violate the constriant c} \\ w_c & \text{if i violates the constriant c} \end{cases}$$ **Equation *40***

Where w_c is the weight value defined for the constraint c. By default, the weight of each constraint is 1.0, corresponding to one Pareto rank in the fitness value.

In the case of a feasible solution of the first Pareto front (rank = 1) and other solution of the same rank that violates one constraint, the resulting fitness of the first individual will be in the range {1, 2}, because the $i_{distance}$ is lower than one and the sum of constraints is zero. In the case of the second individual, the fitness will be in the range {2, 3} as the $i_{distance}$ is lower than one and the sum of constraints is one (assuming default weighting for the constraint). At the time of sorting the population and applying selection operators, the second individual will perform as an individual of Pareto rank 2.

Finally, the weight values of the constraints can be modified by the user according of how strictly the rank of an unfeasible solution must increase. One constraint of low importance can have a w_c value of $w_c = 1.0$ or lower, while another constraint with high importance can have a w_c value of $w_c = 2.0$ or higher.

3.6.2 The molecule individual

To reach the target in a CAMD problem via genetic algorithms is necessary to represent chemical compounds as individuals of a population under evolution. A molecule individual must be easy to modify without losing the physical-chemical feasibility. This section explains the strategy to encode compounds.

To represent a molecule as a modifiable object capable of being modified by the genetic operators of an evolutionary algorithm, the most compliant form is through a *molecular graph* [80]. A molecular graph represents constitutional components of a molecule as a chemical graph where the vertices (nodes) are the atoms and the edges are the chemical bonds between them.

A variant of molecular graphs are *hydrogen-suppressed graphs*, where the hydrogen nodes are neglected, these graphs are not ambiguous, although these might appear to be so. Figure 9 and Figure 10 are examples of molecular and hydrogen suppressed graphs, respectively.

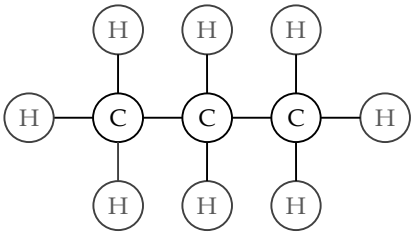


Figure 9. Molecular graph for propane.

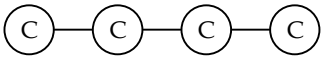


Figure 10. Hydrogen-supressed graph for butane.

This work uses graphs representation for molecules, however the approach selected is more practical and adapted to the properties estimation methods explained above. A molecule individual consists of a graph whose nodes are UNIFAC functional subgroups [81], and the edges are single bonds between those

groups. The CAMD methodology of this work uses the codes for each functional subgroup defined by the UNIFAC Consortium [82], then the mapping of a chemical structure from the subgroups of a graph is done using the cited list of groups. Figure 12 presents the corresponding UNIFAC subgroups graph and the UNIFAC subgroup codes graph for 2-Hydroxybutyric acid (Figure 11).

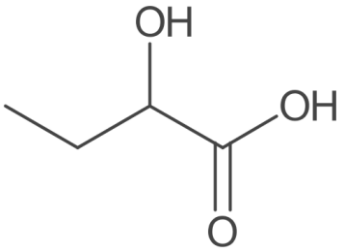
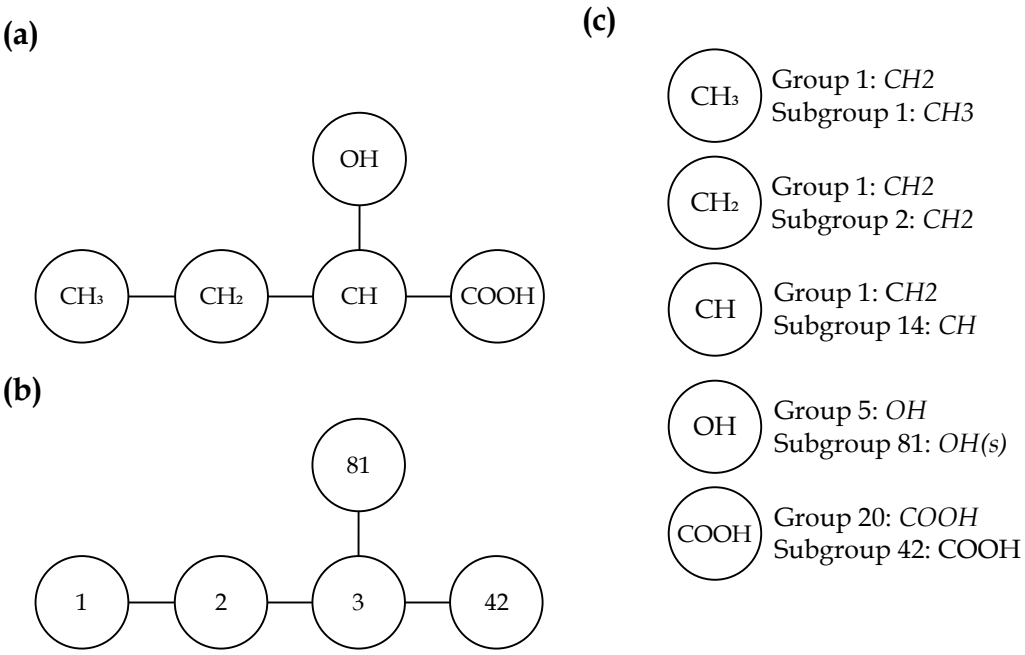


Figure 11. 2-Hydroxybutyric acid, 2D structure.



(a) UNIFAC subgroups graph, (b) UNIFAC subgroups codes graph and (c) Groups and subgroups lists
Figure 12. 2-Hydroxybutyric acid in different representations.

3.6.2.1 Molecule construction rules

In the process of building an individual molecule, the molecule graph formed must meet the following rules:

- The number of bond edges of each functional subgroup node must be equal to the valence of the subgroup. For example, the number of bonds of a CH_3 (1) subgroup must be one, as that is the respective valence, while the subgroup CH=CH (6) must have two bonds.
- The CAMD methodology allows to set the maximum number of functional group nodes per molecule, no individual can contain more functional groups than the allowed number. The default maximum nodes per molecule is ten.
- The number of *strong* subgroup nodes should not exceed a maximum permitted of three. A subgroup is strong if it does not belong to the alkyl or

alkenyl groups. The alkyl group refers to the alkanes of the main group 1 (subgroups 1 to 4) and the alkenyl group refers to the alkenes of the main group 2 (subgroups 5 to 8).

3.6.3 Genetic operators

In an evolutionary algorithm, each offspring is generated by the application of genetic operators over the parent population. The most common operators are single-point crossover and single-point mutation. Both are included in optimization problem of this work and new ones are introduced. The complete list of genetic operators implemented in the CAMD methodology are described below.

3.6.3.1 Mutation

Mutation consists in the selection of one molecule individual from the population, then a subgroup node of the individual is replaced by another subgroup of the same valence selected randomly from the UNIFAC subgroups list. If the number of strong subgroups of the molecule to mutate is equal to the maximum allowed, the selected subgroup is replaced by a not-strong subgroup. Figure 13 is an example of this operator.

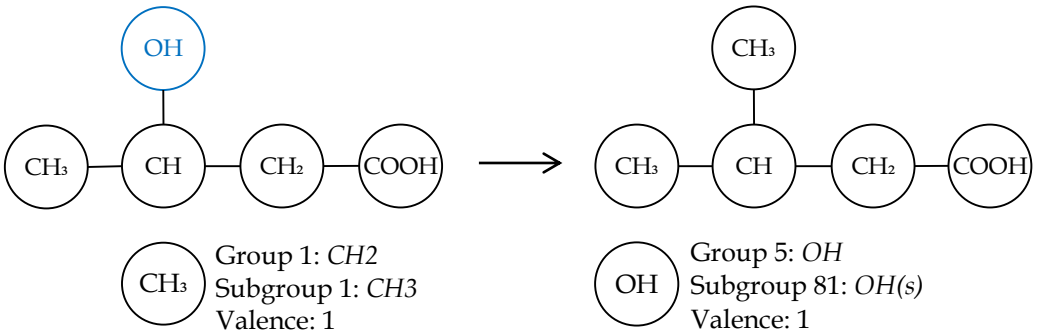


Figure 13. Mutation of a molecule individual.

3.6.3.2 Cross-over

Cross-over consists in the selection of two molecule individuals from the population, then each molecule is split in two parts and those parts are rejoined to produce two different new molecules. One part of the origin molecules is present in each one of the resulting molecules. If any of the resulting molecules is composed only by strong groups, an alkyl group is added between the two parts used to build the molecule.

Figure 14 is an example of this operator.

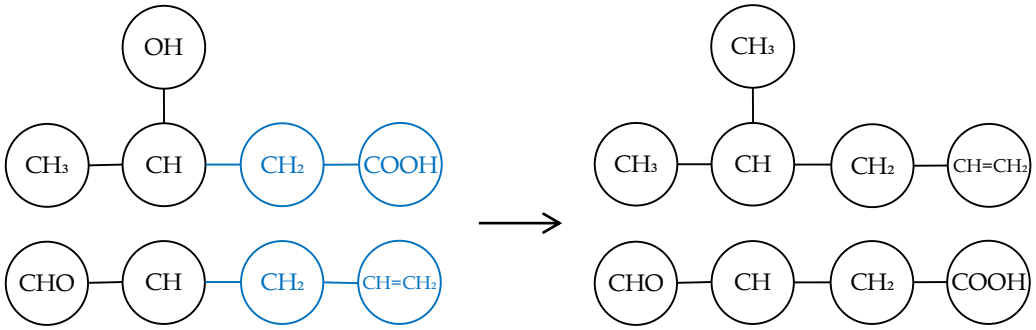


Figure 14. Cross-over of a pair of molecule individuals.

3.6.3.3 Group removal

Group removal consists in the selection of one molecule individual, then an alkyl subgroup node of the individual is selected and one contiguous subgroup of valence one is removed from it. Being *n* the valence of the selected alkyl subgroup, this subgroup is replaced by an alkyl subgroup of valence of *n - 1* in order to meet the rule which states that a subgroup node must be bonded to a number of other nodes equals to its valence. Figure 15 is an example of this operator.

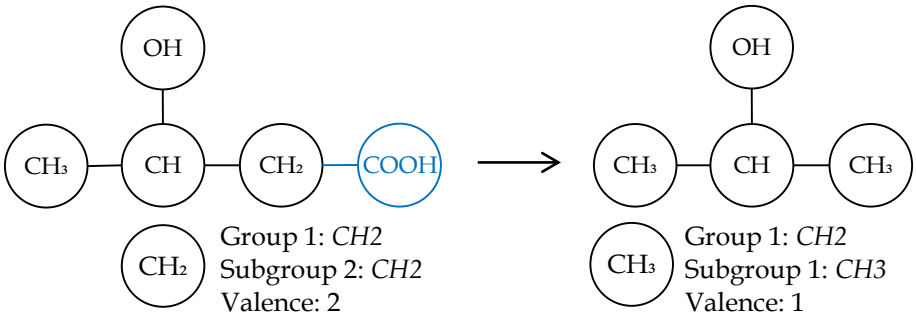


Figure 15. Group removal of a molecule individual.

3.6.3.4 Chain extension

Chain extension consists in the selection of one molecule individual from the population, then the molecule is split in two and then rejoined with an extra CH_2 subgroup between the parts. Figure 16 is an example of this operator.

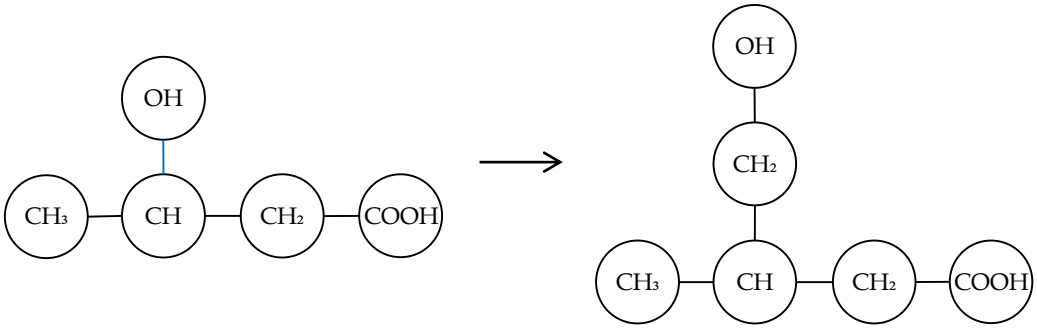


Figure 16. Chain extension of a molecule individual.

3.6.3.5 Chain closure

Chain closure consists in the selection of one molecule individual from the population, the molecule is split in two, one of the parts is discarded and the other is joined with a subgroup randomly selected from the UNIFAC subgroups list. If

the number of strong subgroups of the resulting molecule is equal to the maximum allowed, the selected subgroup is replaced by an alkyl CH_3 subgroup. Figure 17 is an example of this operator.

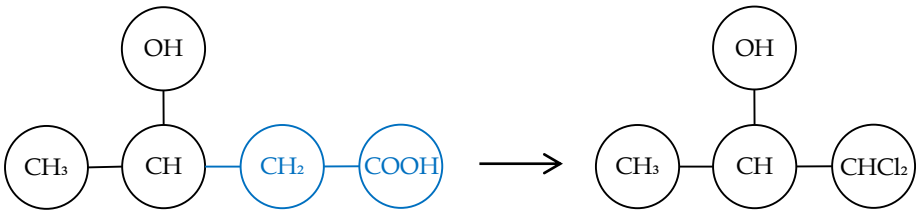


Figure 17. Chain closure of a molecule individual.

3.6.3.6 Chain replacement

Chain replacement consists in the selection of one molecule individual from the population, another molecule is randomly generated according to the rules defined in section 3.6.2, and a cross-over operation is applied to both molecules as explained in section 3.6.3.2. One of the resulting individuals is discarded and the other is the result of the operation. Figure 18 is an example of this operator.

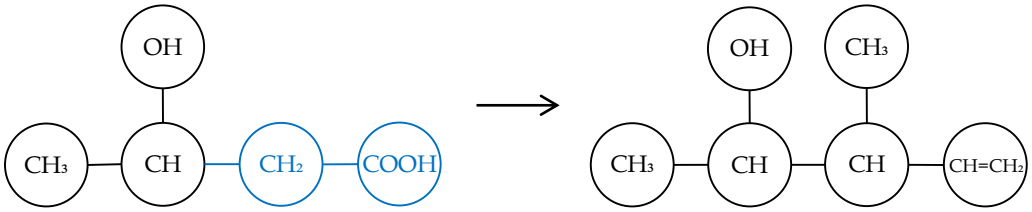


Figure 18. Chain closure of a molecule individual.

3.6.4 Selection mechanism

Each of the genetic operators requires the selection of certain number of individuals from the population. The selection of individuals is done by *Tournament*. In this CAMD methodology, tournament consists in the selection of four molecule individuals from the population, then the individual with best fitness is selected. If a genetic operator requires the selection of n individuals from the population, n tournaments are performed.

3.6.5 Constraints weights

As described in section 3.6.1, constraints handling in the MOHAEA algorithm is based on the constraint-domination approach. Under this approach all the constraints have a weight of 1.0 by default. Table 4 contains the weights of the constraints presented in section 3.1.5.2.

Table 4. Constraint weights.

Constraint name	Weight
Melting point	1.0
Boiling point	1.0
Standard Gibbs energy of formation	1.0

Solvent loss[5]	1.0
Market Availability	1.0

3.7 IMPLEMENTATION OF THE CAMD METHODOLOGY

The thermo-physical, the environmental-related and the mixture properties estimation methods are implemented in java as well as the MOHAEA algorithm, the genetic operators and the interface of the CAMD program. In the case of market availability, the compounds used to query candidate solvents are stored in a MongoDB instance. The source code of the CAMD program is stored in a public repository of github [83]. Anyone is free to view, pull and execute the program.

3.7.1 Program input

The user interface of the program allows the introduction of the next parameters.

- Compound to separate (solute problem)
- Solvent to separate from (solvent problem)
- Separation temperature
- Number of generations for the evolution
- Number of individuals of the first generation
- Number of strong groups
- Functional group families to include in the molecule individuals

3.7.1.1 Molecule input

The molecules introduced to the program must meet the following rules.

- Atoms of the molecule are not directly introduced. UNIFAC-Dortmund subgroups as these are defined by the DDBST GmbH [74] must be introduced. For example, the representation of water is '*16*' as it is the UNIFAC-Do subgroup for this compound.
- A dot '.' represents the single bond joining two subgroups. Double and triple bonds must be represented as the UNIFAC-Do subgroups defined for that. The representation for ethanol is '*1.2.14*', as the numbers represent the CH_3, the CH_2 and the OH groups, respectively. Propylene is represented as '5.1', being 5 the number for the CH_2=CH subgroup
- For branched subgroups, a substructure in parenthesis indicates that it corresponds to a branch emerging from the previous subgroup. In the case of isobutanol the configuration is '*1.3(1).2.14*'. The *(1)* substructure is a branch emerging from of the subgroup 3.

3.7.2 Program output

Once the program has designed a set of optimal solvents, the candidate solvents are displayed in the left side bar of the program window. In this side bar, the structure of each compound is shown and below a table with the next properties.

[5] In Section 4.3 of results, the impact of this constraint on the solvents designed by the CAMD methodology is discussed and the constraint weight is redefined.

- Market availability
- Smiles representation
- Molecule configuration (UNIFAC groups)
- Yield KS
- Environmental Index
- Number of violated constraints
- Number of individuals with the same structure in the end of the last generation.
- Molecular weight.

In the path *output/* respect to the location of the program the next files are created:

- A file with the detailed thermo-physical and environment-related properties of the individuals of the last generation
- Images of the structure of each molecule individual of the last generation.
- Images with the Pareto best front for each generation.

4 Results and discussion

This chapter presents the results of the proposed CAMD methodology for the design of optimal solvents for the separation of lactic acid from an aqueous solution under different conditions. Firstly, results of the property estimation methods for preselected datasets of compounds are presented. Then, results of executions of the program at different conditions are presented and analysed. Finally, to remark the improvements this work, results for the proposed CAMD methodology are compared to the results obtained in the methodology of Serrato [4].

4.1 PROPERTIES ESTIMATION

In the evaluation of the property estimation methods, different datasets of reference compounds were used according to the type of properties. Thermo-physical, environment-related and mixture properties are the types of properties present in this work. Each dataset consists of a list of representative compounds with experimental values of the target properties.

4.1.1 Thermo-physical properties

Accuracy of the estimation methods for these properties is evaluated against a dataset of 108 compounds. Appendix 1 presents the experimental values and the computed values for the dataset. The compounds are a selection of the most common chemical groups distributed as follows.

Table 5. Distribution of the thermo-physical properties dataset.

Chemical group	Number of Compounds
Alkane	12
Alkene	3
Alcohol	24
Aldehyde	5
Ketone	11
Esther	8
Amines	6
Amides	2
Nitrile	3
Ether	6
Carboxylic acids	19
Alkyl halide	9

For the compounds in the dataset, the experimental values for the properties normal melting point, normal boiling point and density are obtained using the Toxicity Estimation Software Tool (T.E.S.T. [84]) developed by the United States Environmental Protection Agency, EPA. The sources of this software tool are the EPI Suite of the U.S. EPA [85]. In the case of standard Gibbs energy of formation, the data used is the contained in the work of Serrato and the Perry's Chemical Engineers' Handbook [86].

4.1.1.1 Normal melting point T_m

For the compounds of the dataset, experimental values of this property and values computed using the Hukkerikar contribution-group method [60], adopted by the CAMD methodology of this work, are represented in Figure 19.

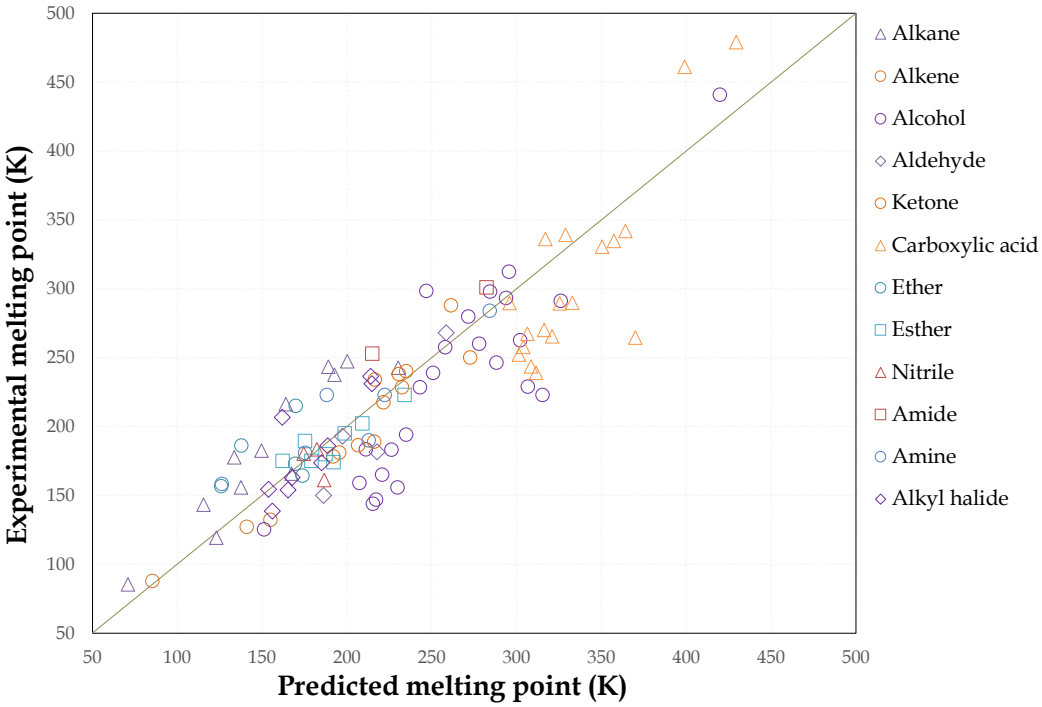


Figure 19. Experimental and predicted values for normal melting point.

In the figure, the data points are distributed in a region around the diagonal, that is an indicator that the prediction model works. The carboxylic acid is the chemical group with the greatest number of compounds distant from the diagonal, while esther compounds lay very close to the diagonal. The indicator used to calculate the error is the Average Relative Error (ARE), which is defined as the mean value of data values' relative error [87]. In Equation *41*, X_j^{exp} and X_j^{pred} are the experimental and predicted values of the element $j = \{1,2,\dots,n\}$.

$$ARE = \frac{1}{n}\sum_{j}\frac{\left|X_j^{exp} - X_j^{pred}\right|}{X_j^{exp}} \times 100 \qquad \textbf{Equation } \textit{41}$$

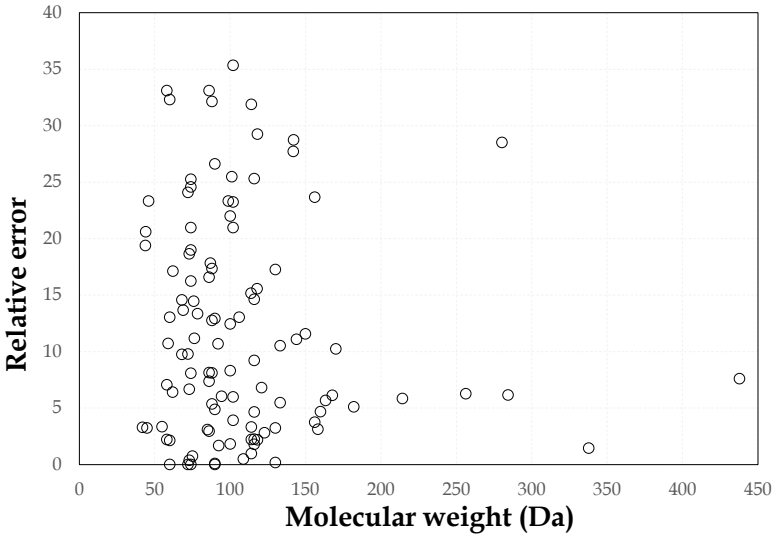


Figure 20. Relative error for melting point.

The average relative error of the predicted data is of 12.12%, very far from the 5.07% that literature reports for the normal melting point method of Hukkerikar. A reason for the regular performance in the prediction of this property can be related to the dependency of melting point of intermolecular interactions and molecular symmetry [88], and as group contribution methods do not consider stereometric factors, providing a very good estimation is difficult. Figure 20 shows that the error is normally higher at low molecular weight, this behavior occurs as most of the compounds with high molecular weight are linear, then molecular shape does not have a strong influence on this property.

4.1.1.2 Normal boiling point T_b

For the compounds of the dataset, experimental values of this property and values computed using the Hukkerikar contribution-group method [60] are presented in Figure 21. The figure shows a good estimation for normal boiling point as the major part of data points lay on the diagonal. The average relative error of the predicted data is of 3.13%, a value above the 1.44% reported in literature for the normal boiling point method of Hukkerikar, that value is still very good for a contribution groups method.

Some of the alcohol compounds' points are distant from the diagonal, the experimental boiling point of these compounds is higher than the predicted. The alcohols with this feature correspond to compounds with more than one hydroxyl group in their structure. Generally, alcohol molecules have high boiling point as these form strong hydrogen bonds in liquid phase and the more hydroxyl groups a compound contains, the higher the boiling point is. Hence a reason for the behavior of this compounds can be that the contribution groups method used to predict this property does not address well the effect of hydrogen bonds caused by the presence of more than one hydroxyl group in a molecule.

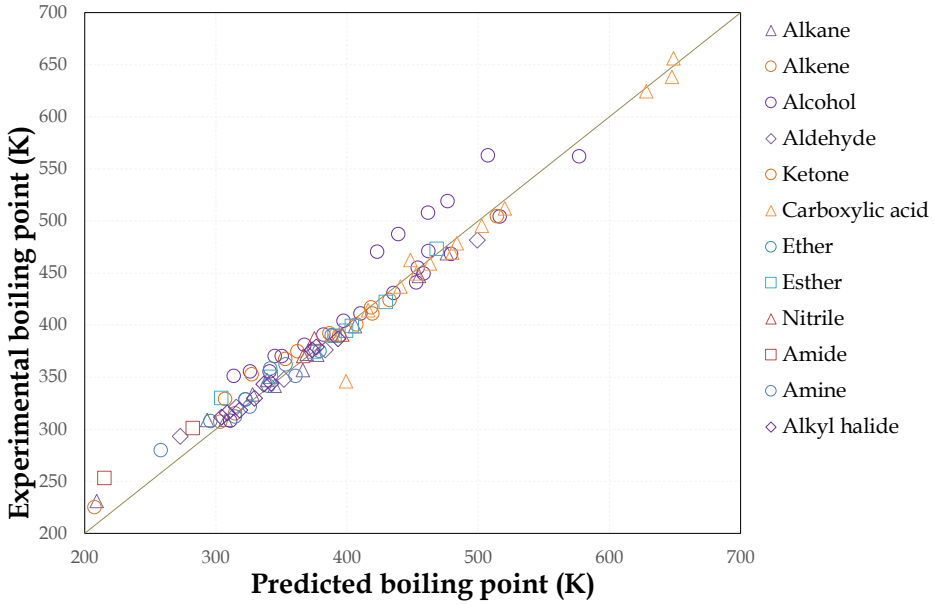


Figure 21. Experimental and predicted values for normal melting point.

4.1.1.3 Standard Gibbs energy of formation, G_f

For the compounds of the dataset, experimental values of this property and values computed using the Hukkerikar contribution-group method [60] are presented in Figure 22. The figure shows an excellent estimation for Standard Gibbs energy of formation as almost all data points lay on the diagonal.

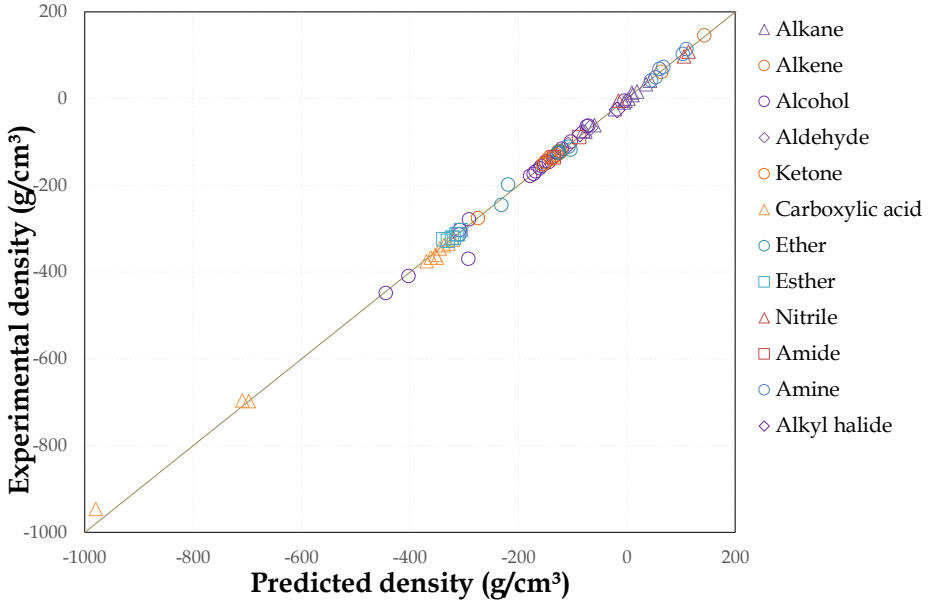


Figure 22. Experimental and predicted values for standard Gibbs energy of formation.

One alcohol and one carboxylic acid are the only compounds that presents values outside the diagonal and anyway these are still close to that line. This property is based in a relative scale, hence applying the average relative error metric is not correct in this case. Then the metric used for standard Gibbs energy of formation is the Average Absolute Error (AAE). In Equation 42, X_j^{exp} and X_j^{pred} are the experimental and predicted values of the element $j = \{1,2,\ldots,n\}$.

$$AAE = \frac{1}{n}\sum_{j}\left|X_j^{exp} - X_j^{pred}\right| \times 100$$ **Equation 42**

The average absolute error of the predicted data is of 5.25 kJ/mol, a value that is practically the same as the 5.24 kJ/mol reported in literature for the standard Gibbs energy of formation of Hukkerikar. This result is impressive, due to the value in literature corresponds to the error for the dataset used to generate the prediction model.

4.1.1.4 Density *ρ*

For the compounds of the dataset, experimental values of this property and values computed using the Hukkerikar contribution-group method [60] are presented in Figure 23. The figure shows a good estimation for density in most of the chemical groups. The average relative error of the predicted data is of 3.92%, a value above the 2.03% reported in literature for the density method of Hukkerikar, that error is still good.

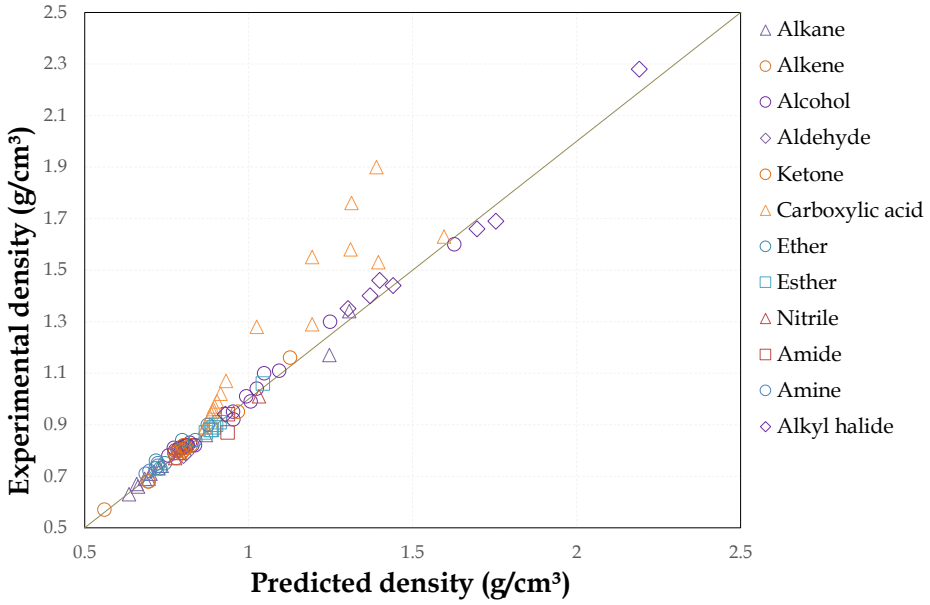


Figure 23. Experimental and predicted values for standard Gibbs energy of formation.

Carboxylic acids is the only group with a significant number of compounds distant from the diagonal. An explanation for this behavior may be related to the formation of strong hydrogen bonds exhibited by carboxylic acids. Hydrogen bonds reduce the intermolecular distance and decrease the molar volume of a compound [89]. Density is the inverse of molar volume.

4.1.2 Environment-related properties

Accuracy of estimation methods for these properties is evaluated against the same dataset of 108 compounds used above for evaluating thermo-physical properties methods. Appendix 2 presents the experimental values and the computed values for the dataset and Table 5 summarizes the distribution of the dataset. These properties were computed as logarithms values.

For the compounds in the dataset, the experimental values for the properties *Fathead Minnow* LC_{50}, *Daphnia Magna* LC_{50}, Oral Rat LD_{50}, Bioconcentration factor and Water solubility were obtained using the Toxicity Estimation Software Tool (T.E.S.T. [84]) developed by the United States Environmental Protection Agency, EPA. The sources of this software tool are listed in the table below.

Table 6. Sources for environment-related properties.

Property	Source
Fathead Minnow LC_{50}	ECOTOX aquatic toxicity database [90]
Daphnia Magna LC_{50}	
Oral Rat LD_{50}	ChemIDplus database [91]
Bioconcentration factor	Arnot, J. A. [92]; EURAS [93]
Water solubility	EPI Suite [85]

Due to the lack of experimental data for many of the compounds of the dataset, the missing values were replaced with predictions done by the T.E.S.T software using the *consensus* strategy which consists in the average of five QSAR prediction methods implemented in this tool [94]. These methods are: Hierarchical method, FDA method, Single-model method, Nearest neighbour method and Group contribution method.

4.1.2.1 *Fathead Minnow* LC_{50}^{FM}

For the compounds of the dataset, experimental values of this property and values computed using the Hukkerikar contribution-group method, implemented in the CAMD methodology of this work, are presented in Figure 24. In the figure, the data points are distributed in a region around the diagonal and there is no remarkable feature related to any chemical group. The absolute average error for the data is of 0.57, relatively close to the 0.48 value reported in literature for the Hukkerikar estimation method [61].

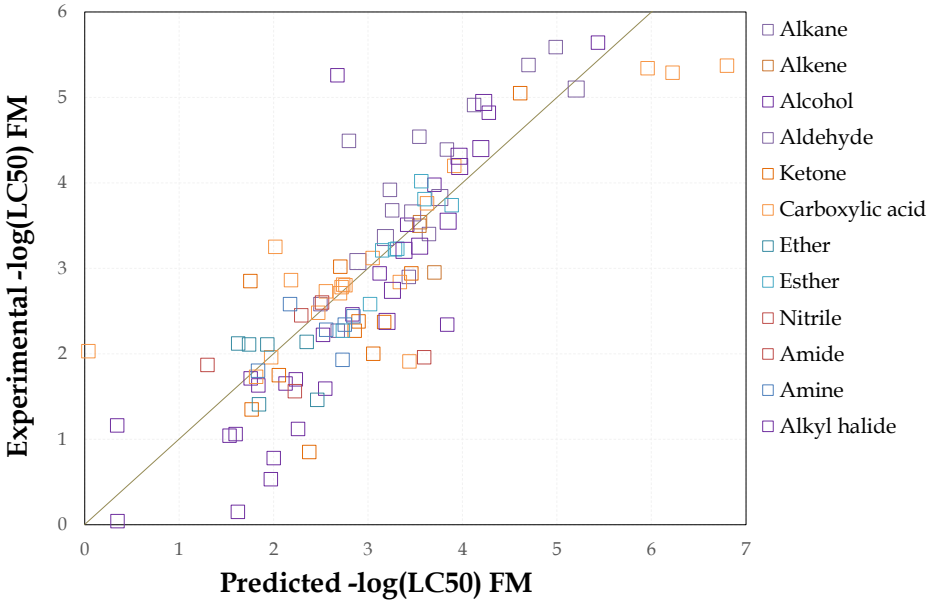


Figure 24. Experimental and predicted values for *Fathead Minnow* LC50.

4.1.2.2 *Daphnia Magna* LC_{50}^{DM}

For the compounds of the dataset, experimental values of this property and values computed using the Hukkerikar contribution-group method are presented in Figure 25. In the figure, the data points are more scattered in comparison with the LC50 *Fathead Minnow* plot. There are no remarkable features related to any chemical group. The absolute average error for the data is of 0.84, a very distant value respect to the 0.49 value reported in literature for the Hukkerikar estimation method [61].

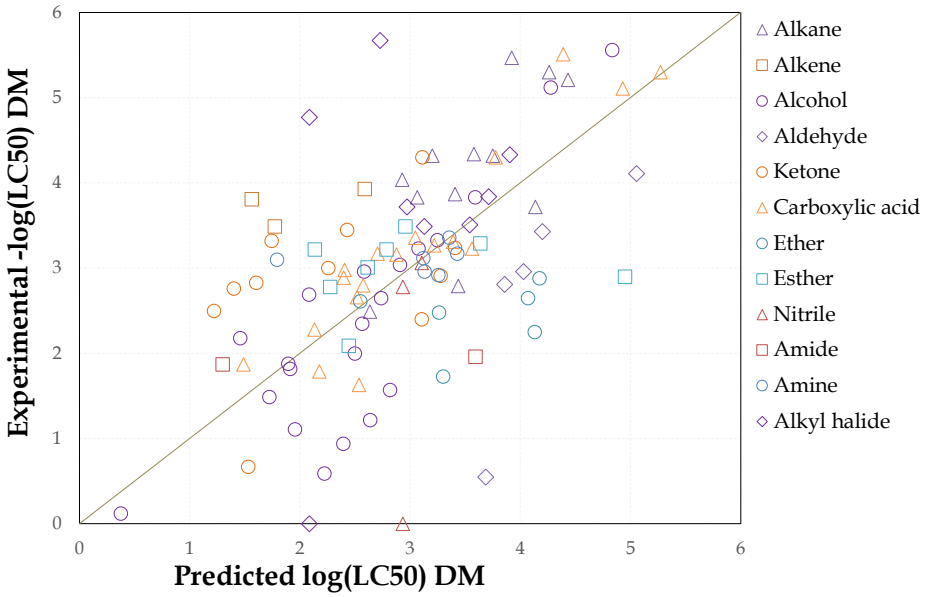


Figure 25. Experimental and predicted values for LC50 *Daphnia Magna*.

4.1.2.3 Oral Rat LD_{50}

For the compounds of the dataset, experimental values of this property and values computed using the Hukkerikar contribution-group method are presented in Figure 26.

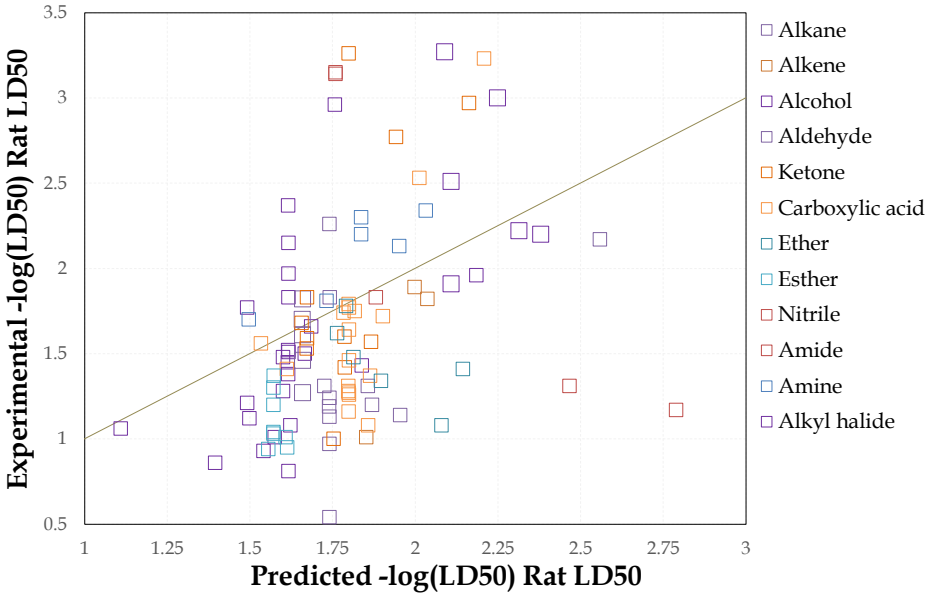


Figure 26. Experimental and predicted values for Rat LD50.

Data points in the figure look scattered. There is no remarkable feature related to any chemical group and the absolute average error for the data is of 0.49, which is not very far respect to the 0.35 value reported in literature for the Hukkerikar estimation method [61]. According to literature, the Oral Rat LD_{50} prediction model is fairly good [39], however that is hard to conclude from the Figure 26.

4.1.2.4 Water solubility W_s

For the compounds of the dataset, experimental values of this property and values computed using the Hukkerikar contribution-group method are presented in Figure 27. In the figure, most of the data points lay over the diagonal line. The compounds whose prediction are far from the diagonal belong mainly to the alkane chemical group. This method is reliable for predicting water solubility of non-alkane compounds, UNIFAC method presents the same problem prediction this property. The absolute average error for the data is of 0.86, not very far from the 0.71 value reported in literature for the Hukkerikar estimation method for this property [61].

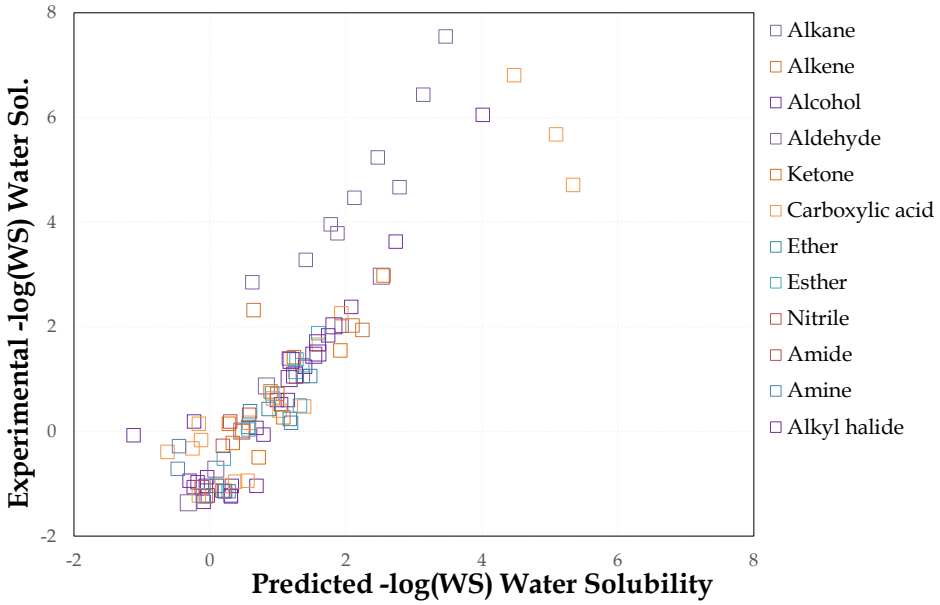


Figure 27. Experimental and predicted values for Water Solubility.

4.1.2.5 Bioconcentration factor BCF

For the compounds of the dataset, experimental values of this property and values computed using the Hukkerikar contribution-group method are presented in Figure 28. In the figure, the data points are distributed in a region around the diagonal and the trend for most of the chemical groups is to have all the compounds laying slightly below or slightly above the diagonal. The absolute average error for the data is of 0.60, above the 0.44 value reported in literature for the Hukkerikar estimation method [61].

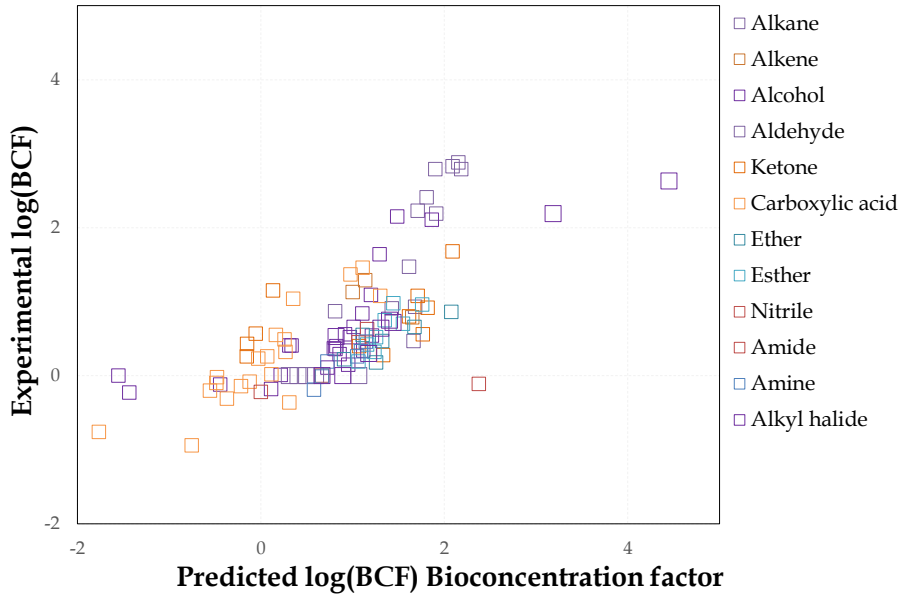


Figure 28. Experimental and predicted values for Bioconcentration factor.

4.1.3 Environmental Index parameters

Now that the estimation methods for environment-related properties are properly described. This section describes the procedure followed to obtain the parameters of the Environmental Index equation in section 3.1.5.1.

The solvents designed using the CAMD methodology of this work must count with advantageous values for the environment-related properties listed in section 3.1.3. The multi-objective optimization approach followed consists in the definition of an optimization function comprising those properties, the first shape of that equation is:

$$E = k_{WS} \log W_s + k_{BCF} \log BFC - k_{LD50} \log LD_{50} - k_{LC50FM} \log LC_{50}^{FM} - k_{LC50DM} \log LC_{50}^{DM}$$

Equation *43*

Where k_{LC50FM}, k_{LC50DM}, k_{WS}, k_{BCF} and k_{LD50} stand for adjustment coefficients for *Fathead Minnow* LC_{50}, *Daphnia Magna* LC_{50}, Oral Rat LD_{50}, Bioconcentration factor and Water solubility, respectively. Five coefficients to be obtained.

A total of 50 000 random compounds are generated following the rules described in section 3.6.2.1 and environment-related properties are computed for each one. The properties of all the compounds are reported as logarithm values. Statistical indicators of each property are presented in the next table.

Table 7. Statistical indicators of environment-related properties of 50 000 compounds.

Property	Mean	Standard deviation	Min	Max
Fathead Minnow LC_{50} (log)	-3.726	1.395	-10.762	0.307
Daphnia Magna LC_{50} (log)	-3.167	1.328	-12.882	2.078
Oral Rat LD_{50} (log)	-2.287	0.520	-5.037	-0.920
Water solubility (log)	-3.848	2.141	-25.179	3.459
Bioconcentration factor (log)	0.938	0.983	-3.664	5.577

Assignment of values for the adjustment coefficients is based on Z-score normalization [95], this technique is used to transform a dataset into a new one with a mean of zero and a standard deviation of one. In this case, this method is used to make the data comparable. A simple way to use Z-score for the assignment of the five adjustment coefficients of Equation *43* is by setting each coefficient as the inverse of the standard deviation of the respective property. This makes possible to control the differences in scale among the properties. However, taking into account that *Fathead Minnow* LC_{50} and *Daphnia Magna* LC_{50} represent a similar environmental issue, which is the impact of a chemical compound on the mortality of aquatic life, the reduction of the scale of both properties by two is a reasonable measure. That is done by dividing the coefficients of the properties by two. Furthermore, both properties have very similar values for standard deviation. Based on the mentioned considerations, the coefficients can be obtained as follows:

$$k_{LC50FM} = \frac{1}{2}\frac{1}{SD_{LC50FM}} \approx \frac{1}{SD_{LC50FM} + SD_{LC50DM}}$$
$$k_{LC50DM} = \frac{1}{2}\frac{1}{SD_{LC50DM}} \approx \frac{1}{SD_{LC50FM} + SD_{LC50DM}} \quad \textbf{Equation 44}$$

$$SD_{LC50FM} \approx SD_{LC50DM} \qquad k_{LD50} = \frac{1}{SD_{LD50}} \qquad k_{WS} = \frac{1}{SD_{WS}} \qquad k_{BCF} = \frac{1}{SD_{BCF}}$$

The number of coefficients can be reduced if we have in mind that in this optimization problem the scale of the objective function is not relevant, then the next transformations can be performed.

$$k_{LC50FM} = 1$$
$$k_{LC50DM} = 1$$
$$k_{LD50} = \frac{SD_{LC50FM} + SD_{LC50DM}}{SD_{LD50}}$$
$$k_{WS} = \frac{SD_{LC50FM} + SD_{LC50DM}}{SD_{WS}} \quad \textbf{Equation 45}$$
$$k_{BCF} = \frac{SD_{LC50FM} + SD_{LC50DM}}{SD_{BCF}}$$

The results are only three coefficients, below Equation *43* is transformed into Equation 14 and the coefficients can be computed.

$$E = k_{WS}\log W_s + k_{BCF}\log BFC - k_{LD50}\log LD_{50} - \log LC_{50}^{FM} - \log LC_{50}^{DM} \quad \textbf{Equation 14}$$

Finally, the values of the adjustment coefficients are the following.

$$k_{LD50} = 5.235 \qquad k_{WS} = 1.272 \qquad k_{BCF} = 2.771 \quad \textbf{Equation 46}$$

4.1.4 Mixture properties

The mixture properties in this work comprises only Infinite Dilution Activity Coefficients in Aqueous Solution, γ^{∞}. Accuracy of the estimation method for this property is evaluated against a dataset of 111 compounds whose experimental values for activity coefficient were obtained from the work of Serrato [4] and the values reported by Mitchell [96]. Appendix 3 present the experimental values and the computed values for the dataset and Table 5 summarizes the distribution.

Table 8. Compounds distribution in mixture properties dataset.

Chemical group	Number of Compounds
Alkane	13
Alkene	8
Alcohol	32
Aldehyde	8
Ketone	12
Esther	8
Amine	8
Amide	2
Nitrile	5
Ether	10
Carboxylic acid	4
Sulfide	1

For the compounds of the dataset, Figure 29 shows the natural logarithm of this property ($\ln \gamma^{\infty}$) applied to experimental values and predicted values computed using the UNIFAC Dortmund method implemented in the CAMD methodology of this work.

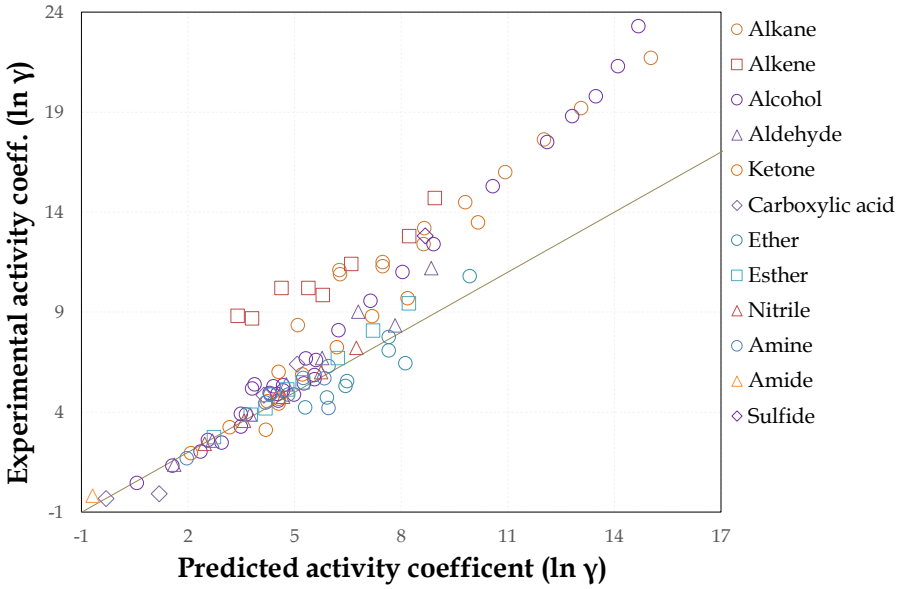


Figure 29. Experimental and predicted values for activity coefficient.

In the figure, the compounds of most of the chemical groups lay close to the diagonal. Nevertheless, a significant number of compounds are distant from the

diagonal, these belong to alkane, alkene and alcohol groups. The poor performance of UNIFAC predicting activity coefficients at infinite dilution in water for alkanes and alkenes is widely known [97].

Figure 30 shows the absolute error against the molecular weight of the compounds. It can be noted that error increases with molecular weight. The reason for this behavior is that larger compounds tend to be similar to alkanes. In the case of alcohol group, this is the group with more compounds in the dataset (Table 8) and the largest compounds of the dataset belong to it.

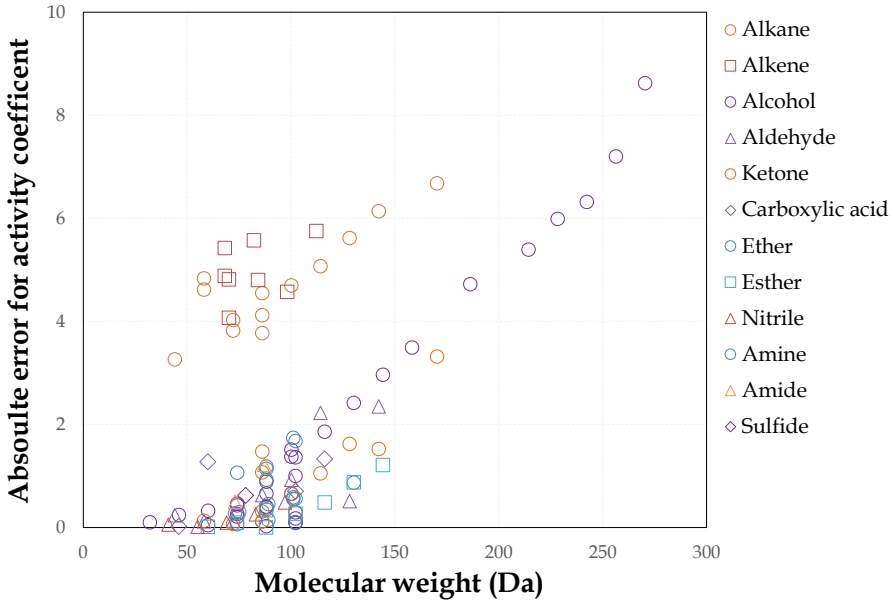


Figure 30. Absolute error for activity coefficient.

4.2 MOLECULES EVOLUTION

The evaluation of the evolution process consisted in inspecting the values of the objective functions and the genetic operator rates along the evolution. For the initial experiments, the generation of results is done by executing 50 times the CAMD program introducing the same separation problem and the same evolution parameters. Table 9 summarizes the configuration of the program executions. In addition to the table, in the construction of molecules and the application of genetic operators, all the functional groups are used except aromatic, sulphur, cyclic, heterocyclic and aromatics.

Table 9. CAMD program initial execution details.

Number of executions	50
Solute to separate	Lactic acid
Problem solvent	Water
Separation temperature	298K
Number of individuals per generation	50
Number of generations	50
Max functional groups per molecule	10
Max active strong per molecule (not alkane)	3

4.2.1 Convergence

The convergence of the MOHAEA algorithm implemented in the CAMD methodology of this work is evaluated by extracting the molecule individual's values for the objectives Yield KS and Environmental Index, for each generation along the evolution in 50 executions of the CAMD program using the configuration mentioned above. The average and the standard deviation of the objective values of every generation are computed for the executions. Figure 31 and Figure 32 show the progress of the average and the standard deviation of the optimization objectives values.

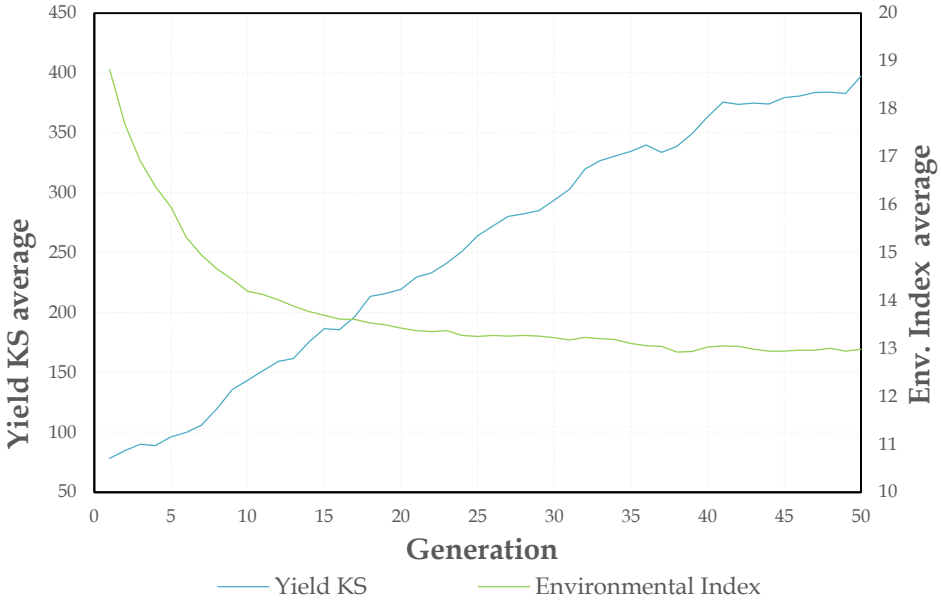


Figure 31. Average for Yield KS and Environmental Index.

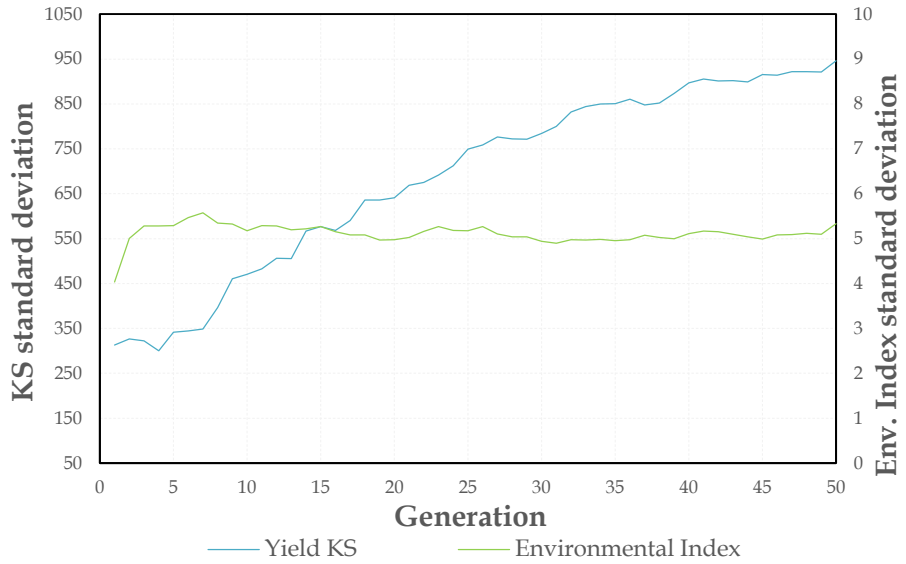


Figure 32. Standard deviation for Yield KS and Environmental Index.

In Figure 31, the average of the optimization objectives are computed to inspect how population obtained, generation by generation, better values for the objectives. In the case of Environmental Index, 50 generations are enough to converge

to an optimal average value. On the other hand, in the case of Yield KS, 50 generations are not enough to converge to an optimal value as in all generation there is an increase in the average of this objective function.

In Figure 32, the standard deviations for the optimization objectives are computed generation by generation to inspect in what extent the solutions were spread over the solutions space. In the case of Environmental Index, a convergence point is reached by the same generation the average does. In the case of Yield KS, there is an increase along all generations.

For the purpose of finding a convergence point for the KS objective, a new experiment using the same configuration of the last experiment is performed with 150 generations instead of 50. Figure 33 and Figure 34 contain the progress of the average and the standard deviation of the optimization objectives values.

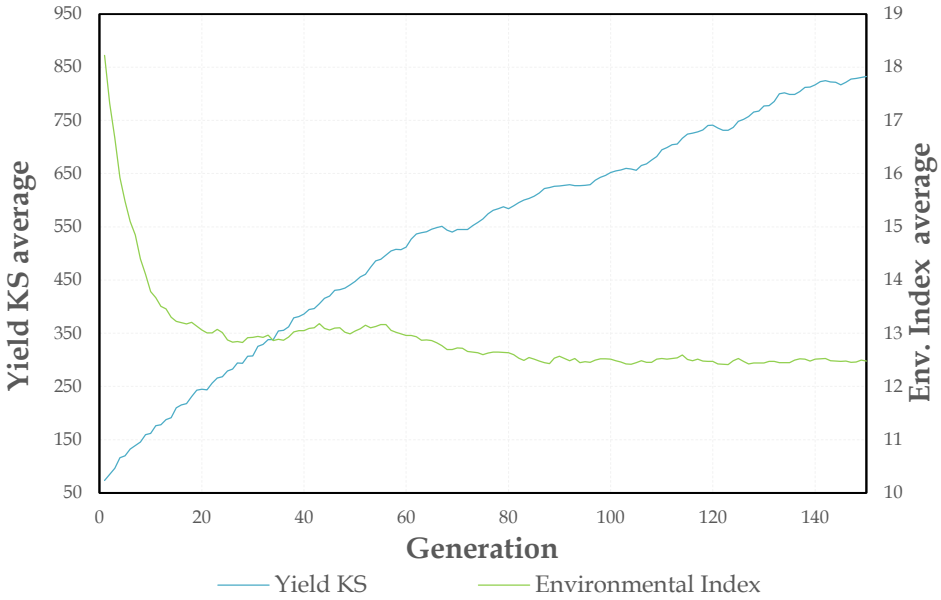


Figure 33. Average for Yield KS and Environmental Index, 150 generations.

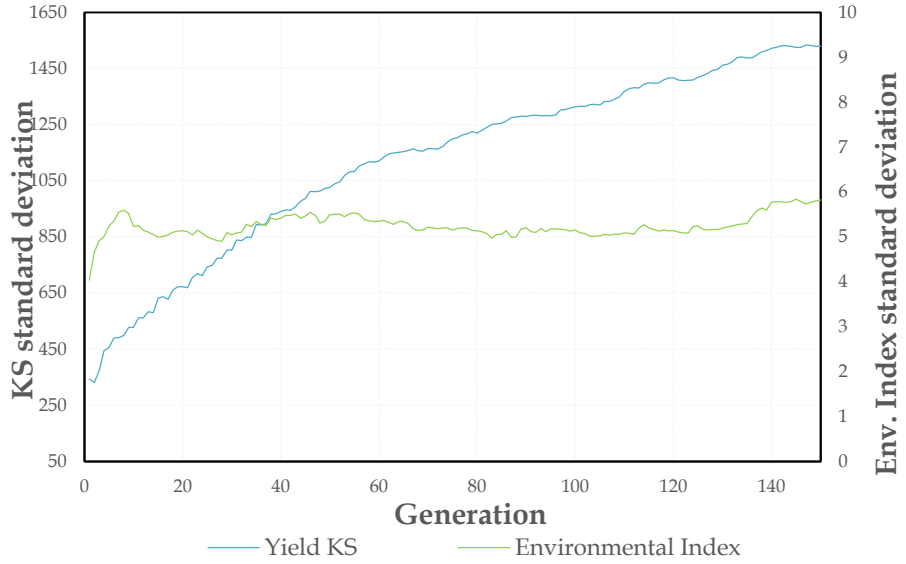


Figure 34. Standard deviation for Yield KS and Environmental Index, 150 generations.

Both figures show that after 150 generations Yield KS does not reach a convergence point yet, the graph presents only a slight reduction in the growth rate. This behavior indicates that there is no convergence point for Yield KS and later analysis in Section 4.3.1 will show that the solutions with the greatest values for this objective do not meet the constraints of the problem and the solutions meeting all the constraints lay in ranges of very low values. Even so, 150 generations is the reference for the analyses of the next sections.

A final experiment, under the same conditions of the last one, is conducted in order to inspect the evolution of the solutions that meet all the constraints. In this experiment, only the values of the objective functions of feasible solutions are considered. Figure 35 shows the progress of the average of the optimization objectives values. It can be pointed out how the scale of the Yield KS is reduced radically and in the case of the Environmental Index, the scale remains practically untouched. The behavior of both objectives indicates that the constraints of the problem influence strongly the evolution of Yield KS while the affectation over the evolution of the Environmental Index is minimal.

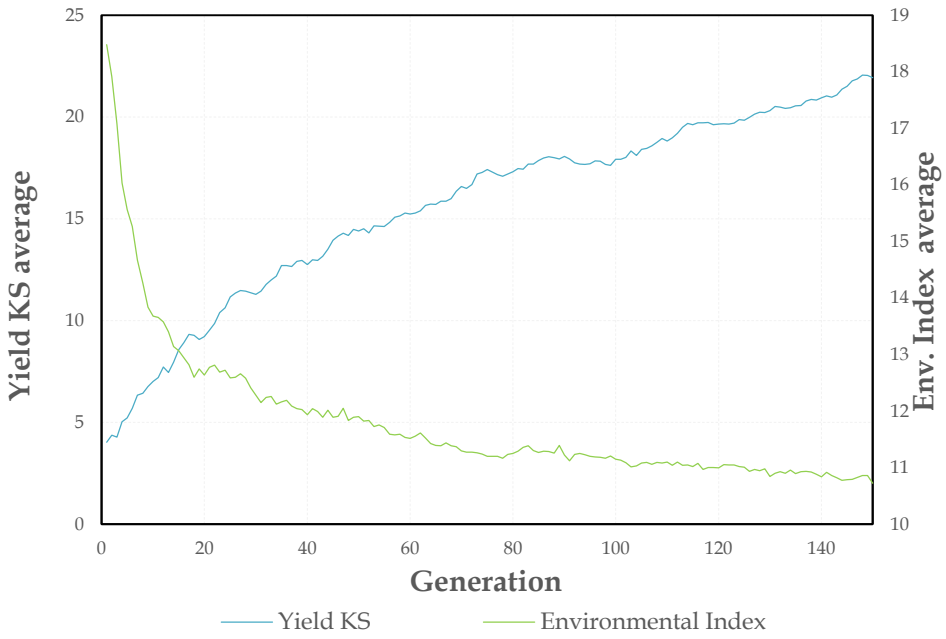


Figure 35. Average for Yield KS and Environmental Index of feasible solutions, 150 generations.

4.2.2 Operators performance

The six genetic operators implemented in the CAMD methodology of this work are evaluated by extracting the application rates of the genetic operator of each generation along the evolution in 50 executions of the CAMD program using the configuration mentioned above. The average of the application rates of each operator, each generation, are computed for those executions.

Figure 36 shows the average of the application rates of the genetic operators, generation by generation, due to the noise in the results, the experiment is performed again with 150 executions. Figure 37 contains the results of the new experiment, it is the average of the application rates of the genetic operators, generation by generation. In the figure, noise is reduced, but it is still present.

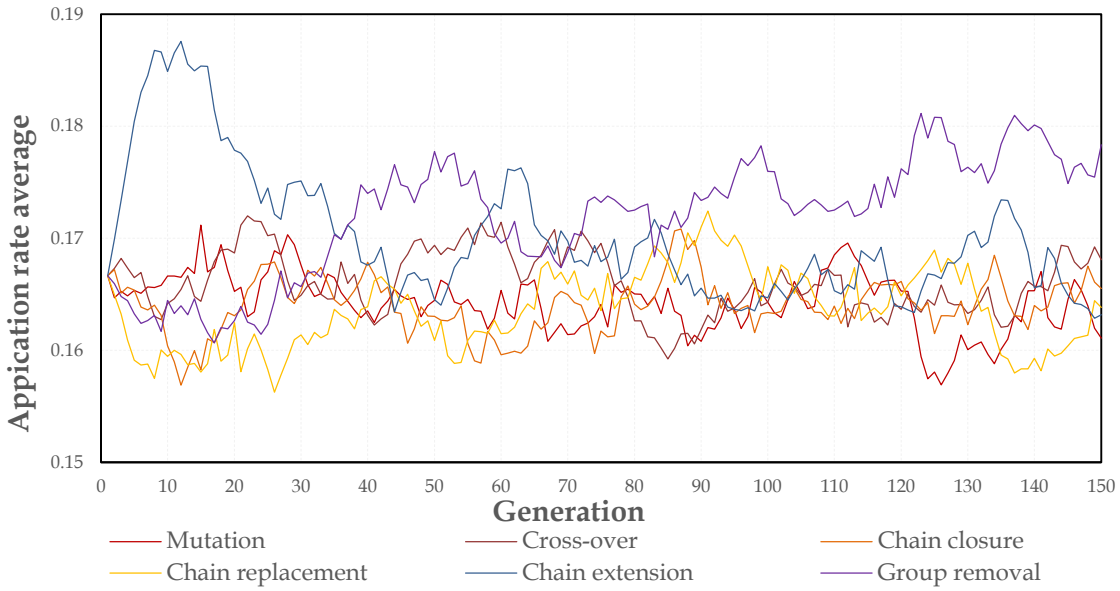


Figure 36. Average of genetic operator rates along the evolution

The main features of Figure 37 are as follows. In the early stages of the evolution, the most outstanding operator is *Chain extension*, whose rate increases in the first generations while the rates of the other operators decrease. Chain extension peaks in the 10th generation, then a decline starts towards a convergence point. In the case of *Group removal,* in the early stages the application rate of this operator decreases and around the 15th generation it rises towards the highest convergence point. In the case of *Chain closure* and *Mutation*, the behaviour of both is the same, the rates of these operators decline slightly along the evolution with some fluctuations, and toward the lowest convergence points. In the case of *Cross-over*, this operator fluctuates along the evolution and there are no meaningful peaks in their progress. In the case of the *Chain replacement* operator, the rate of this operator decline strongly in the first stages of the evolution and then it increases slightly till a low convergence point.

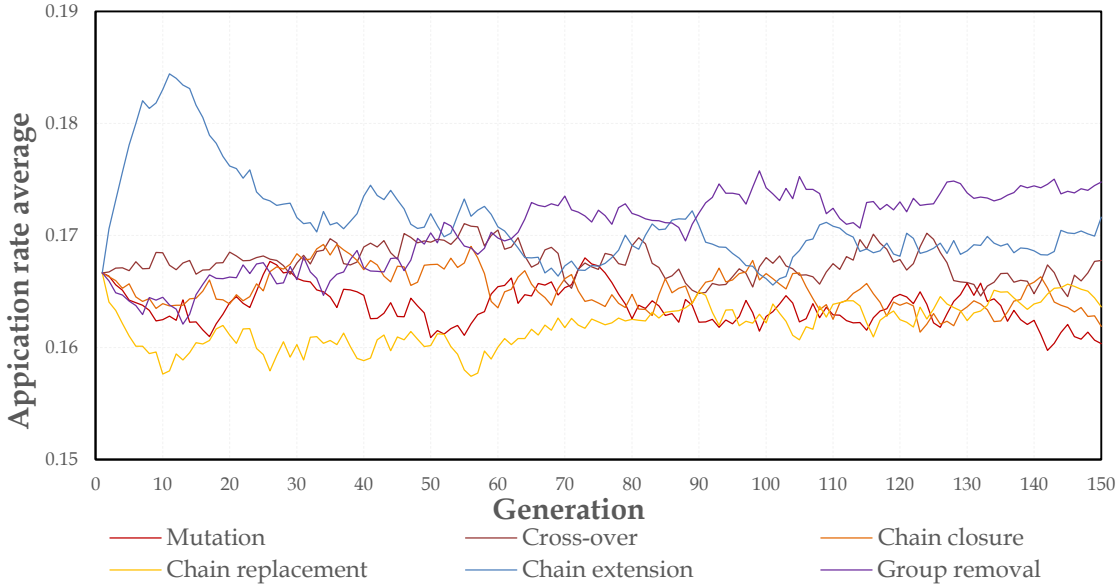


Figure 37. Average of genetic operator rates along the evolution

The behaviours described above can be explained as follows. On the whole, the number of generations is not enough to present a common converge point for the operators. As it mentioned in Section 4.2.1, the Yield KS objective function does not converge in 150 generations. At the early stages, Mutation and Group removal present a bad performance due to these operators are intended to promote exploitation, and during the first generations the promotion of exploration is more important as the initial pool of solutions are not expected to be located near any optimal.

The Chain replacement operator is the only one that presented a poor performance overall, as the rate of this operator remains at the lowest values along the evolution, and the convergence point is one of the lowest. A reason for this behaviour is related to the fact that this is the strongest operators, as once applied it adds and unknown random chain to a molecule individual, causing a hard-to-predict impact over the spread of a molecule parent. The performance of this operator shows that molecule individuals are very sensitive.

In the case of the Chain extension operator, its success is due to this operator is not strong as it changes the size of a molecule in a very controlled way. When a good molecule individual is affected by this operator, the resulting molecules can keep the strong features of the original molecule.

4.3 SOLVENTS DESIGN

The evaluation of the compounds designed by the CAMD methodology of this work consists in running the program several times and analyse the designed solvents taking into account separation power, environmental performance and market availability.

4.3.1 Base experiment

In an initial experiment, the generation of results is done by executing 50 times the CAMD program introducing the separation problem and the evolution parameters summarized in Table 9 in all executions. The result of the first experiment is summarized in Table 10.

Table 10. CAMD program results summary, first experiment.

Best solvents designed	318
Unique best solvents designed	157
Solvents meeting all constraints	12
Unique solvents meeting all constraints	6

The table does not show a good performance for the CAMD methodology, as only four compounds meet the constraints of the problem. Table 11 and Figure 38 present the constraints of the problem and the number of compounds that meet the constraints.

Table 11. Constraints in the best solutions, first experiment.

Constraint name	Compounds	Percentage
Melting point	234	73.58%
Boiling point	313	98.43%
Standard Gibbs energy	318	100.00%
Solvent loss	21	6.60%
Market availability	285	89.62%

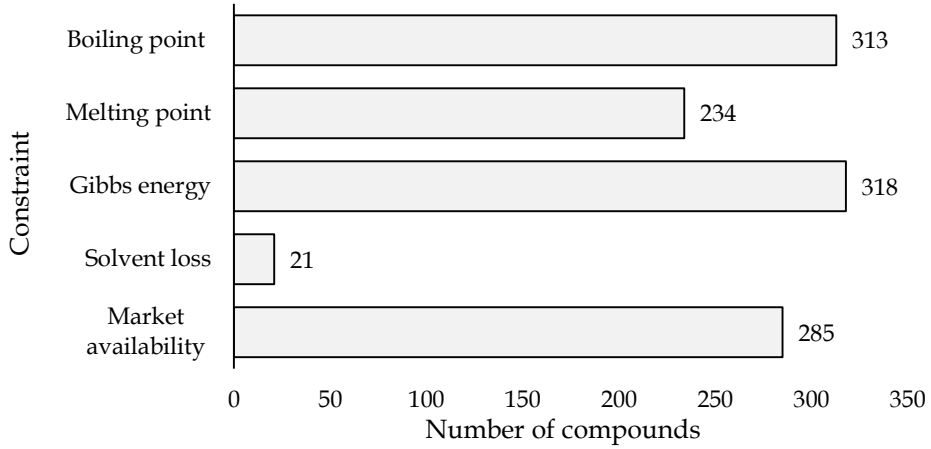


Figure 38. Number of solutions meeting each constraint, first experiment.

Not many compounds meet the Solvent loss constraint, while for the other constraints the number of compounds is acceptable. A reason for the low number of compounds meeting the Solvent loss constraint is due to this property does not depend exclusively of the nature of the compound, but mostly of the interactions between the compound and the problem solvent. These interactions make this property hard to control and the associated constraint hard to meet.

To make that more compounds meet the Solvent loss constraint, the constrain definition is softened and the constraint weight increased. The new constraint is defined in Equation *47* and the new weights of the problem are in the Table 4.

$$L < 0.15 \qquad \textbf{Equation 47}$$

Table 12. Modified constraint weights

Constraint name	Weight
Melting point	1.0
Boiling point	1.0
Standard Gibbs energy of formation	1.0
Solvent loss	2.0
Market Availability	1.0

Using these new parameters, a second experiment is conducted where results are generated again by executing 50 times the CAMD program with the same separation problem and the same evolution parameters presented above. The results of the executions are summarized in Table 13.

Table 13. CAMD program results, second experiment.

Best solvents designed	318
Unique best solvents designed	156
Solvents meeting all constraints	40
Unique solvents meeting all constraints	16

The table shows a significant improvement respect the results of the first experiment, as now 40 solvents meet the constraints of the problem. Table 14 and Figure 39 present the individual improvement for each constraint. In this case, a larger number of compounds meet the Solvent loss constraint, while the number of compounds meeting the other constraints slightly decreases in all cases.

Table 14. Constraints in the best solutions, second experiment.

Constraint name	Compounds	Percentage
Melting point	230	72.33%
Boiling point	310	97.48%
Standard Gibbs energy	318	100.00%
Solvent loss	74	23.27%
Market availability	272	85.53%

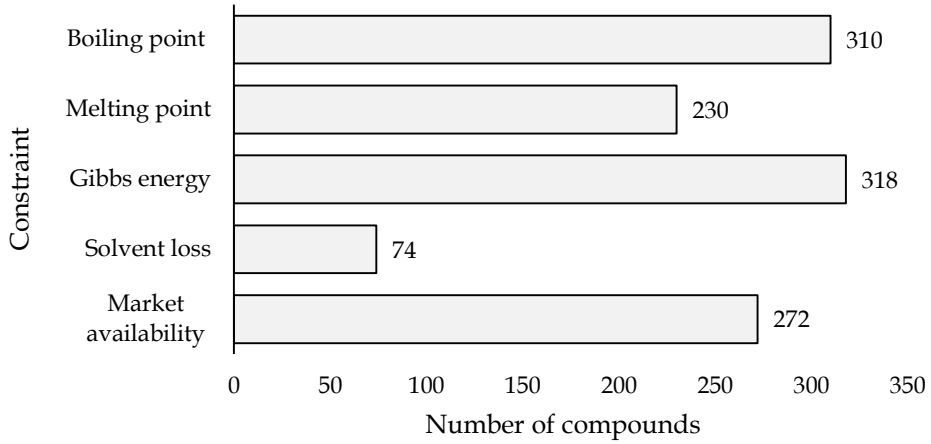


Figure 39. Number of solutions meeting each constraint, second experiment.

In line with the results presented above, Table 12 and Table 15 summarize the definitive configuration used in the executions of the CAMD program performed for the rest of this chapter.

Table 15. CAMD program definitive execution summary.

Number of executions	100
Solute to separate	Lactic acid
Problem solvent	Water
Separation temperature	298K
Number of individuals per generation	50
Number of generations	150
Max functional groups per molecule	10
Max strong groups per molecule (not alkane)	3

Using these definitive configuration, a third experiment is conducted. Results are generated again and a summary of the executions is presented in Table 15. The table shows that in this experiment better solvents are designed and there are more solvents meeting all constraints compared to the two previous experiment, however the results of the experiments are not comparable as the number of executions and generations is different in this experiment. This experiment has been named as the *base experiment* for the rest of this chapter.

Table 16. CAMD program execution summary, third experiment.

Best solvents designed	668
Unique best solvents designed	276
Solvents meeting all constraints	71
Unique solvents meeting all constraints	28

Table 17 and Figure 40 present the percentage of solvents meeting each constraint. The percentages for this experiment are similar to the reported for the second experiment in Table 14.

Table 17. Constraints in the best solutions, third experiment.

Constraint name	Compounds	Percentage
Melting point	495	74.10%
Boiling point	641	95.96%
Standard Gibbs energy	667	99.85%
Solvent loss	169	25.30%
Market availability	538	80.54%

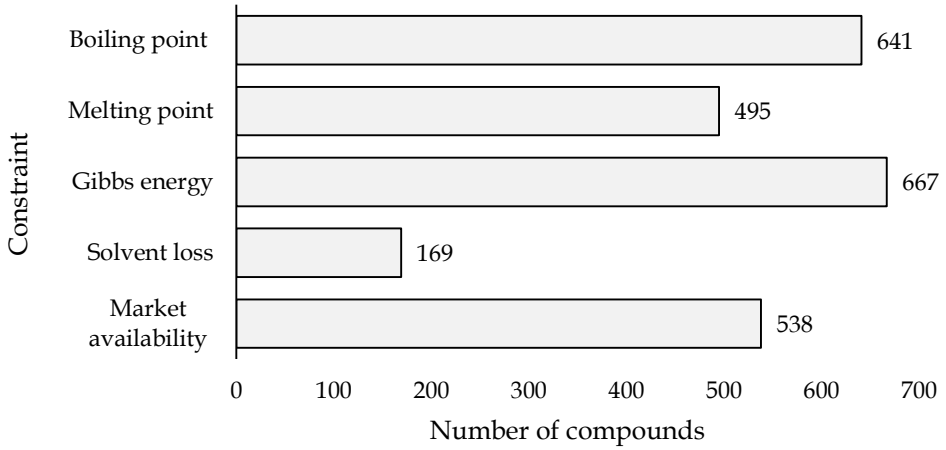


Figure 40. Number of solutions meeting each constraint, third experiment.

The solutions that appear the most in the best solvent sets produced by the executions are the shown in Figure 41. Carbonyl compounds are reluctant to be part of the best solvents set and only two compounds meet the problem constraints, these are the *4-Oxobutyl acetate* (SMILES: CC(=O)OCCCC=O) and the *3,3-Dichloropropanal* (SMILES: C(C=O)C(Cl)Cl). The rest of the compounds do not meet the solvent loss constraint.

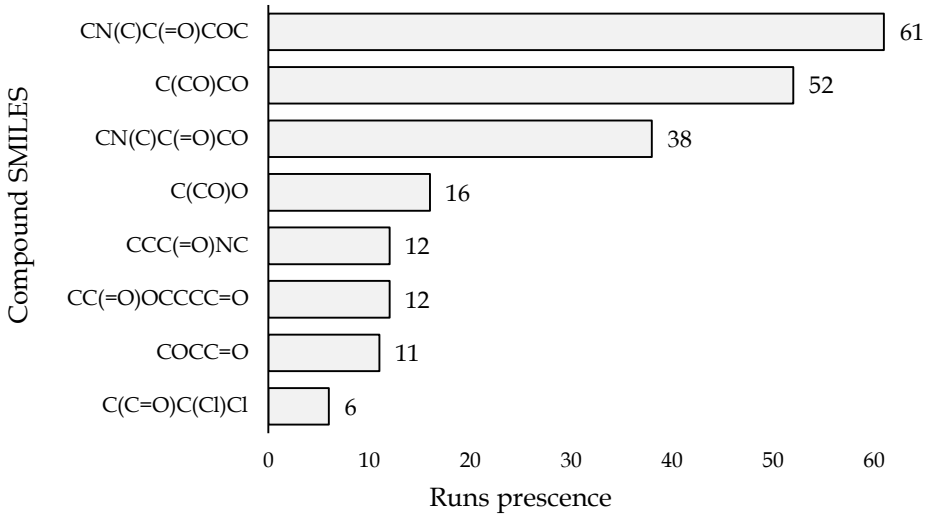


Figure 41. Most frequent compounds in the executions.

The selection of the best solutions among the 276 unique candidates is done by the identification of the Pareto best front. Figure 42 shows the resulting distribution of solutions and the best front. In the figure, the shape of the Pareto best front indicates that effectively a trade-off exists between the solvent power and the environmental impact of a solvent. In addition, for the decision makers the selection of a good solvent is not an easy task, as the best designed solvents are distributed in a concave Pareto front. The issue of multi-objective problems with concave Pareto fronts is that the best solutions tend to be located in the extremes [98].

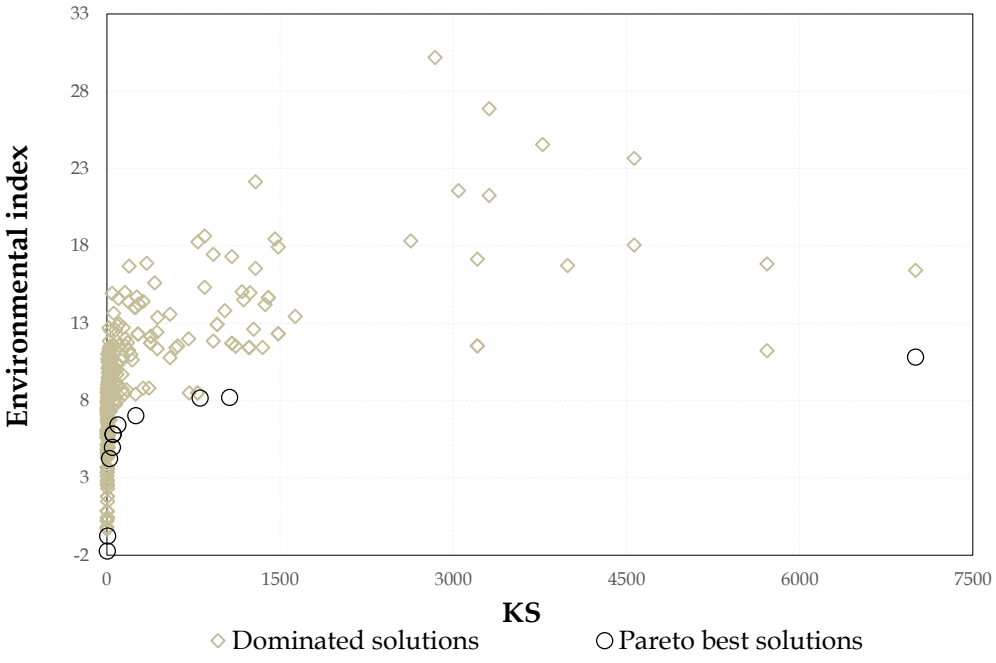


Figure 42. Best Pareto front and dominated solutions.

In the figure, the Pareto best front comprises 11 compounds, however none of them meet all the problem constraints. Table 18 summarizes these solvents and

the number of constrains violated by each one. Most of the best compounds violate more than one constraint. An inspection of the constraints violated shows that the most violated constraint is the market availability constraint, as the 11 compounds violate this constraint.

Table 18. Pareto best compounds.

Solvent SMILES	Runs presence	KS	E	violated constraints
C(CC(=O)CCCO)CC(=O)C=O	1	5.9	-0.76	3
C(CC=O)CN(CC=O)C=O	1	7002.8	10.82	2
C(CCC(=O)CC=O)CC=O	2	53.8	5.82	2
C(CCC=O)CCC(=O)C=O	1	53.8	5.82	2
C=C(C)C=CC(=O)CCN(CCC=O)C=O	1	22.7	4.25	3
C=C(CC=CCC(=O)CN(CCC=O)C=O)CC=O	1	94.8	6.43	3
CC(=O)C(=O)CCN(C)C(=O)CC(=O)N	1	250.0	7.04	5
CC(=O)CCCCC(=O)CC(=O)C	1	2.2	-1.73	2
CCC(=O)N(CO)CO	1	47.1	4.99	3
CCN(C)C(=O)CC(=O)N(C)CO	1	807.8	8.16	3
CN(CO)C(=O)COC	1	1061.8	8.21	2

To produce a better set of best compounds, a new set of Pareto best front are obtained using only the designed compounds that met all the constraints. Figure 43 shows the resulting distribution of solutions and the Pareto best front. Table 19 shows the Pareto best compounds.

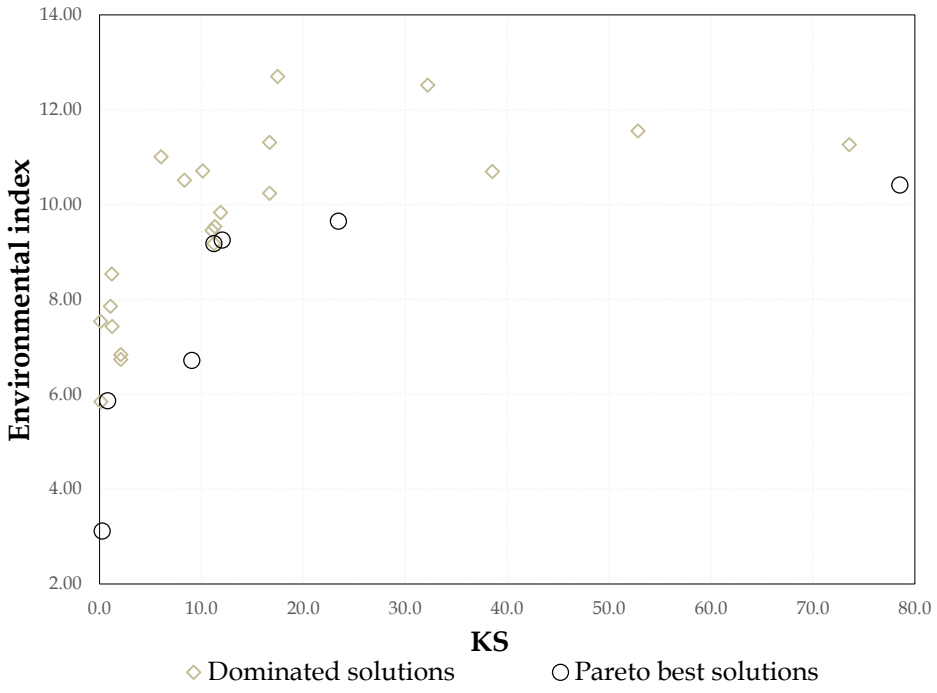


Figure 43. Pareto best front with feasible compounds, solutions.

Table 19. Best feasible solutions.

SMILES	Runs presence	KS	EI
CC(=O)OCCCC=O	12	23.5	9.65
C(C(Cl)Cl)O	7	12.0	9.25
C(CCC=O)CC=O	6	78.5	10.41
CCCC(=O)OO	2	11.2	9.17
CC(=O)CCCOC(=O)C	1	0.8	5.86
CC=CC=CCCC(=O)C	1	0.3	3.11
CCCCCC(=O)C=O	1	9.1	6.71

This time, the resulting Pareto best front comprises 7 Pareto best compounds. Comparing these best *feasible* solvents with the *unfeasible* solvents obtained including all compounds (Figure 42), the major difference is not only the size of the solutions set, but also the range of the objective functions. Table 20 summarizes these ranges.

Table 20. Ranges of Pareto best fronts.

Best compounds	Solvent yield, KS		Environmental index, E	
	min	max	min	max
Unfeasible compounds	2.16	7002.8	-1.73	10.82
Feasible compounds	0.27	78.5	3.11	10.41

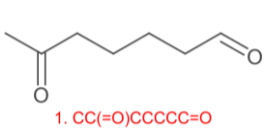
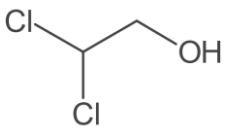
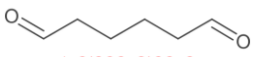
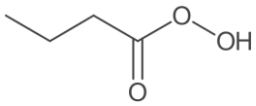
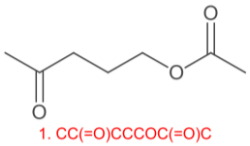
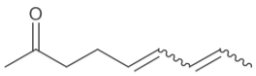
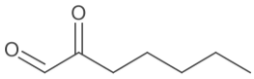
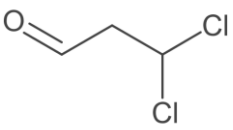
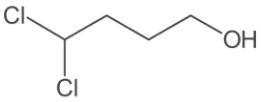
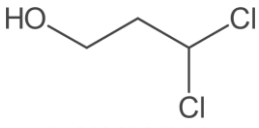
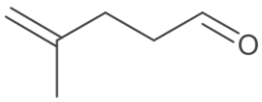
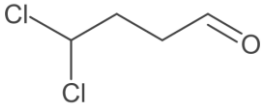
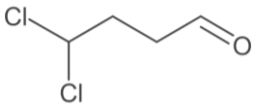
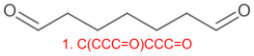
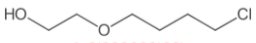
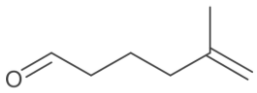
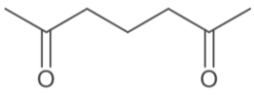
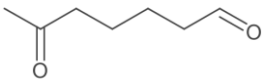
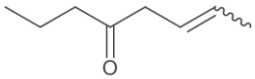
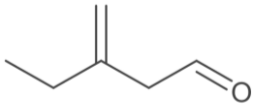
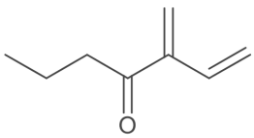
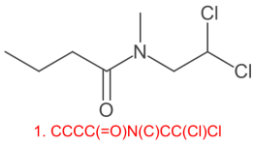
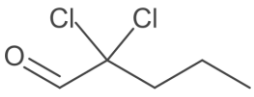
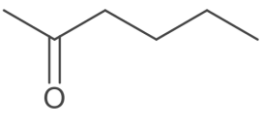
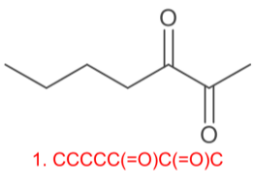
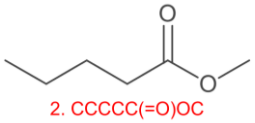
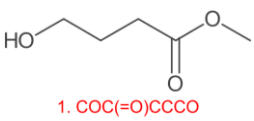
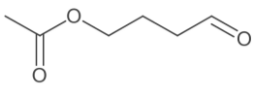
The ranges of the unfeasible Pareto best solvents are much wider compared with the ranges of the feasible best solvents. The large reduction of the range for the Yield KS objective and the continuous increase of this objective along the evolution shown in Figure 33 indicate that this objective does not have an optimal converge point and the constraints of the problem keeps the best feasible solutions remains in stable range, despite of the growth of this objective function caused by unfeasible solutions.

The next sections will inspect the feasible solvents of Figure 43. Table 21 and Table 22 show their chemical identifier and structure, respectively. More detailed tables are presented in the next sections.

Table 21. Best feasible solvents identifiers.

#	CAS	Name	SMILES
1	6564-95-0	4-Oxobutyl acetate	CC(=O)OCCCC=O
2	598-38-9	2,2-Dichloroethan-1-ol	C(C(Cl)Cl)O
3	1072-21-5	Hexanedial	C(CCC=O)CC=O
4	13122-71-9	Butaneperoxoic Acid	CCCC(=O)OO
5	5185-97-7	4-Oxopentyl acetate	CC(=O)CCCOC(=O)C
6	62765-21-3	Nona-5,7-dien-2-one	CC=CC=CCCC(=O)C
7	2363-85-1	2-oxoheptanal	CCCCCC(=O)C=O
8	14538-09-1	3,3-Dichloropropanal	C(C=O)C(Cl)Cl
9	159433-58-6	4,4-dichlorobutan-1-ol	C(CC(Cl)Cl)CO
10	83682-72-8	3,3-Dichloropropan-1-ol	C(CO)C(Cl)Cl
11	3973-43-1	4-methylpent-4-enal	C=C(C)CCC=O
12	101257-10-7	4,4-dichlorobutanal	C(CC(Cl)Cl)C=O
13	7307-02-0	Heptane-2,4-dione	CCCC(=O)CC(=O)C
14	53185-69-6	Heptanedial	C(CCC=O)CCC=O
15	78450-84-7	2-(4-Chlorobutoxy)ethan-1-ol	C(CCOCCO)CCl
16	64825-78-1	5-methylhex-5-enal	C=C(C)CCCC=O
17	13505-34-5	Heptane-2,6-dione	CC(=O)CCCC(=O)C
18	19480-04-7	6-Oxoheptanal	CC(=O)CCCCC=O
19	185678-49-3	Oct-6-en-4-one	CC=CCC(=O)CCC
20	445423-02-9	3-methylidenepentanal	CCC(=C)CC=O
21	50360-62-8	3-methylidenehept-1-en-4-one	CCCC(=O)C(=C)C=C
22	-	N-(2,2-dichloroethyl)-N-methyl-butanamide	CCCC(=O)N(C)CC(Cl)Cl
23	41718-50-7	2,2-Dichloropentanal	CCCC(C=O)(Cl)Cl
24	591-78-6	Hexan-2-one	CCCCC(=O)C
25	96-04-8	Heptane-2,3-dione	CCCCC(=O)C(=O)C
26	624-24-8	Methyl pentanoate	CCCCC(=O)OC
27	925-57-5	Methyl 4-hydroxybutanoate	COC(=O)CCCO
28	6149-41-3	Methyl 3-hydroxypropanoate	COC(=O)CCO

Table 22. Feasible best solvents structures.

1  1. CC(=O)CCCCC=O	2  2. C(C(Cl)Cl)O	3  1. C(CCC=O)CC=O	4  4. CCCC(=O)OO
5  1. CC(=O)CCCOC(=O)C	6  1. CC=CC=CCCC(=O)C	7  2. CCCCCC(=O)C=O	8  1. C(C=O)C(Cl)Cl
9  1. C(CC(Cl)Cl)CO	10  2. C(CO)C(Cl)Cl	11  1. C=C(C)CCC=O	12  1. C(CC(Cl)Cl)C=O
13  2. C(CC(Cl)Cl)C=O	14  1. C(CCC=O)CCC=O	15  2. C(CCOCCO)CCl	16  1. C=C(C)CCCC=O
17  1. CC(=O)CCCC(=O)C	18  1. CC(=O)CCCCC=O	19  1. CC=CCC(=O)CCC	20  1. CCC(=C)CC=O
21  5. CCCC(=O)C(=C)C=C	22  1. CCCC(=O)N(C)CC(Cl)Cl	23  2. CCCC(C=O)(Cl)Cl	24  1. CCCCC(=O)C
25  1. CCCCC(=O)C(=O)C	26  2. CCCCC(=O)OC	27  1. COC(=O)CCCO	28  2. CC(=O)OCCCC=O

4.3.2 Representative solvent candidates

The strategy used in this work to evaluate the designed solvents generated by the CAMD methodology is based on the strategy used by Serrato [4], where the program is executed a certain number of times, the resulting optimal solvents from the executions are merged and the best solvent candidates are reported and inspected. This approach is valid when the goal of the research is the solution of a specific chemical product design problem, however in this work is important to evaluate the quality of the method used to design solvents and ensure that the solution of another problem requires only a few executions of the problem and not the 50 or the 100 executions performed in the last section.

Evaluating the quality of the method requires the extraction and inspection of a set of designed solvents representing a typical single execution of the CAMD program. The extraction is done by extracting the different Pareto dominance fronts from the set of designed solvents obtained in the base experiment (Table 17) and locating the Pareto front that covers the half of the solution solvents. The selection of the representative solutions among the 668 solvent candidates produced a subset of 72 where 22 were unique compounds.

Figure 44 shows the resulting solutions of the experiment, the representative Pareto front and the most frequent solutions. In the figure, the representative front split the solutions by half, hence the frequency of the designed solvents is well distributed along the results of the experiment and there are no regions with accumulation of repeated solvents. Regarding the most frequent designed solvents, Figure 44 displays the solutions with presence in at least 10 executions of the program in the experiment (Figure 41). Such solutions are not concentrated but distributed in the solution space, and these form a line covering a wider range of objective values than the representative Pareto front.

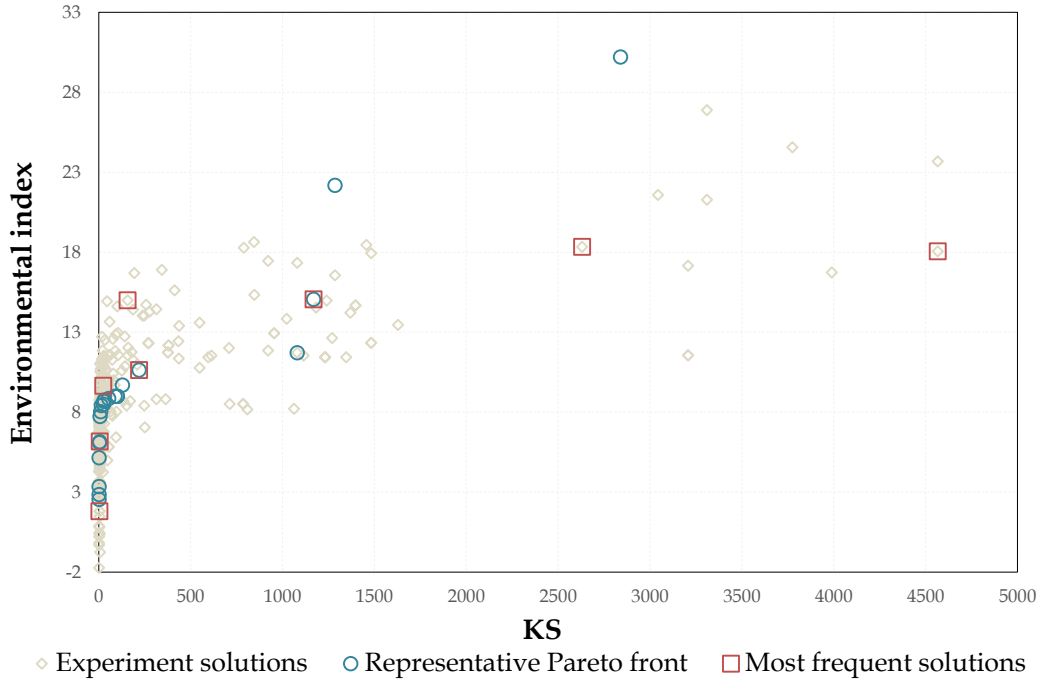


Figure 44. Representative Pareto front and experiment solutions.

4.3.3 Separation of lactic acid from an aqueous solution

The properties of importance to evaluate the separation power of the candidate feasible solvents are tabulated in Table 23. The table comprises the feasible compounds as explained above, hence all the properties meet the solvent requirements.

Table 23. Feasible candidate solvents, thermo-physical properties.

#	Melting Point	Boiling Point	Density	Gibbs Energy	Solvent Loss	KS
1	264.4	457.4	0.939	-416.7	0.124	23.46
2	259.5	415.6	1.234	-199.6	0.128	12.04
3	282.9	463.6	0.865	-202.3	0.141	78.54
4	261.0	432.2	0.925	-458.8	0.148	11.23
5	270.3	464.0	0.908	-442.6	0.066	0.79
6	270.9	469.2	0.776	38.9	0.036	0.27
7	288.4	469.1	0.844	-228.0	0.083	9.07
8	252.7	416.9	1.198	-155.1	0.046	73.56
9	273.0	462.8	1.124	-183.4	0.035	11.88
10	266.4	440.3	1.169	-191.5	0.075	11.31
11	229.2	402.4	0.722	-29.5	0.134	16.69
12	259.9	441.5	1.146	-147.0	0.020	52.84
13	287.7	454.3	0.925	-262.1	0.043	1.21
14	288.8	484.1	0.863	-194.2	0.069	38.55
15	282.3	499.1	0.984	-273.4	0.038	6.06
16	237.7	428.4	0.738	-21.4	0.051	8.35
17	285.1	465.7	0.836	-258.3	0.130	2.08
18	285.4	480.8	0.854	-224.5	0.134	10.13
19	252.2	441.5	0.733	-39.5	0.025	0.15
20	226.6	401.9	0.724	-29.5	0.134	16.69
21	251.8	439.8	0.771	30.8	0.132	1.09
22	255.2	383.6	1.560	-44.6	0.023	32.20
23	291.5	454.0	1.567	-116.2	0.010	17.49
24	225.8	392.4	0.682	-134.1	0.049	0.12
25	285.5	464.7	0.837	-258.1	0.130	2.08
26	249.9	422.8	0.784	-227.9	0.119	1.23
27	267.9	455.3	0.914	-450.7	0.070	11.05
28	261.0	432.2	0.925	-458.8	0.148	11.23

Besides Yield KS, Solvent loss is the most important property of the table. A low value for solvent loss is a desirable feature as in the design of a chemical process, the costs of the inputs as well as the steps and equipment of the separation train are heavily influenced by this property. In the table, there is a high variability among the solvents, the range varies from 2% to 14.6%, covering almost all the feasibility range of the constraint for this property.

With the aim of evaluating the relationship between solvent loss and KS, Figure 45 is a plot this property against Yield KS. In the figure, there are solutions with good values for both yield and solvent loss, therefore there is no visible trade-off between both properties.

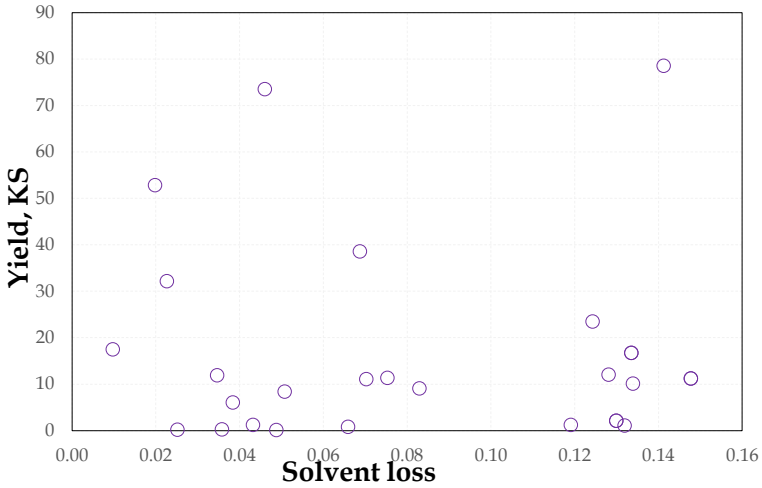


Figure 45. Solvent loss and Yield.

4.3.4 Environmental performance of candidate solvents

The properties of importance to evaluate the environment performance of the feasible candidate solvents are tabulated in Table 24. Reviewing the ranges of the properties in the table, water solubility is the property with the largest range, hence the selection of solvents with very low water solubility is possible. This is advantageous not only for the environment but also for the reduction of solvent loss in the case that the solute to separate is in aqueous solution, the impact of this reduction in a process is discussed above.

Table 24. Feasible candidate solvents, environment-related properties.

#	log(LC50FM)	log(LC50DM)	log(LD50)	log(WS)	log(BFC)	Env. Index, E
1	-4.10	-3.58	-1.51	-5.38	0.33	9.65
2	-2.44	-2.59	-1.87	-5.40	0.48	9.25
3	-4.55	-5.16	-1.60	-5.31	-0.33	10.41
4	-2.67	-3.70	-1.63	-5.18	0.30	9.17
5	-2.68	-2.83	-1.64	-5.97	-0.23	5.86
6	-2.43	-1.85	-1.93	-8.16	-0.33	3.11
7	-3.21	-4.20	-1.81	-6.04	-0.89	6.71
8	-3.65	-4.22	-1.91	-5.53	0.15	11.26
9	-3.02	-2.93	-1.87	-6.08	0.66	9.83
10	-2.73	-2.76	-1.87	-5.74	0.57	9.53
11	-4.00	-3.21	-1.70	-5.92	0.60	10.24
12	-3.94	-4.39	-1.91	-5.87	0.25	11.55
13	-1.50	-3.28	-1.91	-1.56	-1.54	8.53
14	-4.84	-5.33	-1.60	-5.65	-0.24	10.69
15	-3.14	-2.84	-2.10	-6.08	0.64	11.00
16	-4.29	-3.38	-1.70	-6.28	0.70	10.51
17	-2.33	-2.09	-1.82	-5.75	0.04	6.73
18	-3.96	-3.22	-1.61	-5.36	0.69	10.71
19	-2.79	-1.60	-1.87	-7.15	0.28	5.84
20	-3.78	-4.75	-1.75	-5.88	0.39	11.31
21	-3.22	-4.30	-2.24	-7.64	-0.60	7.85
22	-3.44	-3.60	-1.88	-6.19	1.27	12.52
23	-4.22	-4.03	-1.49	-5.58	1.35	12.70
24	-2.34	-3.45	-1.81	-6.15	0.04	7.54
25	-2.40	-1.87	-1.90	-5.89	0.04	6.83
26	-1.80	-4.07	-1.91	-5.82	-0.38	7.43
27	-2.96	-3.88	-1.63	-5.53	0.40	9.45
28	-2.67	-3.70	-1.63	-5.18	0.30	9.17

To identify relations among the properties, a correlation matrix is made using the five properties and the Environment Index E objective function. Table 25 shows this matrix. The near-zero values in most of the non-diagonal items in the matrix show that there is no strong linear relation among the environment-related properties or the environment objective function.

The correlation with a nearest value to the unit is the correlation between *Fathead Minnow* $\log LC_{50}^{FM}$ and *Daphnia Magna* $\log LC_{50}^{DM}$ with a correlation value of 0.569. This correlation can be expected, as both properties measure the same environmental issue, the impact of compounds on aquatic biota.

Table 25. Correlation matrix, environment-related properties.

	log(LC50FM)	log(LC50DM)	log(LD50)	log(WS)	log(BFC)	E
log(LC50FM)	1.000					
log(LC50DM)	0.569	1.000				
log(LD50)	-0.452	-0.270	1.000			
log(WS)	0.172	-0.202	0.277	1.000		
log(BFC)	-0.397	0.117	0.284	-0.272	1.000	
E	-0.633	-0.568	0.266	0.315	0.574	1.000

4.3.5 Market availability

The impact of market availability on the results of the CAMD methodology is evaluated by comparing the best solvents obtained in the base experiment to the best solvents obtained from experiments excluding the Market Availability constraint. The best solvents of the base experiment are the shown in Figure 43 obtained by the configuration shown in Table 15 where all constraints are included (Table 12). On the other hand, the other best solvents are obtained by an experiment consisting in executing the CAMD program with the configuration of Table 15 and the constraint weights shown in Table 26. The results of this experiment are summarized in Table 27.

Table 26. Constraint weights, no market availability.

Constraint name	Weight
Melting point	1.0
Boiling point	1.0
Standard Gibbs energy of formation	1.0
Solvent loss	2.0
Market Availability	0.0

Table 27. CAMD without Market Availability program execution summary.

Best solvents designed	699
Unique best solvents designed	370
Solvents meeting the constraints	74
Unique solvents meeting the constraints	40
Solvents meeting the constraints and market available	33
Unique solvents meeting the constraints and market available	12

Comparing the results of this experiment with the results of the base experiment shown in Table 16, the number of best Pareto solvents meeting the problem constraints slightly increases from 70 to 74, and the number of unique best solvents meeting the problem constraints significantly increases, changing from 28 to 40.

Inspecting the market availability of the best solvents of this experiment, the number of solvents meeting the problem constraints and available in the market is 33. In this set, 12 solvents are unique, which is a very small number compared with the 28 unique feasible best solvents of the base experiment. Table 28 summarizes those feasible solvents for the experiment.

Table 28. Pareto best solvents, experiment.

SMILES	KS	E	Market availability
CC(=O)OCCC=CCC=O	16.0	8.95	no
CC(=O)OC=CCCCC=O	16.0	8.95	no
C=C(C=CCOC(=O)C)OC(=O)C	2.0	6.58	no
CC=CC(=C)CCC=O	12.1	7.87	no
CC=CCN(CCCC=O)C=O	216.7	10.40	no
CC(=O)OCCCOOC(=O)C	39.9	9.13	no
CCCCCC(=O)C=O	9.1	6.71	yes
CCCCCN(CC=O)C=O	270.6	12.31	no
CCCCN(CCC=O)C=O	270.6	12.31	no
CCCN(CCCC=O)C=O	270.6	12.31	no
CCN(CCCCC=O)C=O	270.6	12.31	no

Figure 46 shows the best feasible solvents of this experiment and the best feasible solvents of the base experiment. In the figure, the range of the Yield KS objective in the Pareto best front for this experiment is wider compared to the range of the same best solvents for the base experiment. On the other hand, the range of the Environmental Index (E) objective in the Pareto best front of the base experiment is wider compared to the range of the best solvents of this experiment.

On the whole, the best solvents resulting from this experiment have higher values for KS and slightly lower value for Environmental Index. In addition, most of the solvents generated in this experiment are not available in the market, and the only one available is one of the few solvents with a good value for the Environmental Index in this experiment.

These results suggest that here is a trade-off between solvent power and market availability that does not exist between environment performance and market availability. For the solvents of this experiment located in the middle of the Pareto best front, these are not far from the Pareto best front of the base experiment and there is one common solvent in the fronts of both experiments, it is the 2-oxoheptanal (SMILES: CCCCCC(=O)C=O).

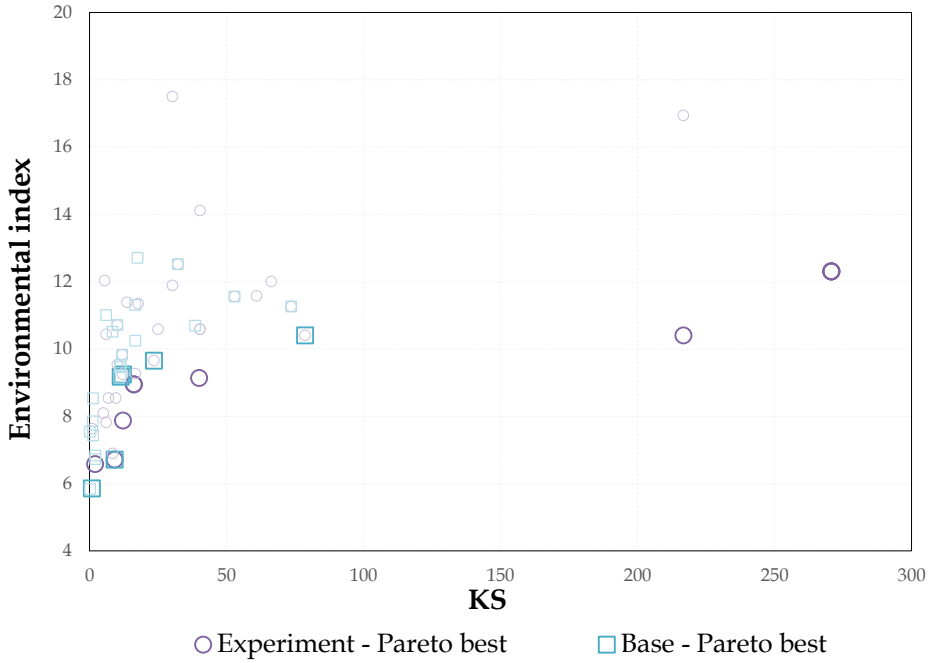


Figure 46. Pareto feasible solutions for the base experiment and the no-availability experiment.

4.4 THE PREVIOUS AND THE NEW METHODOLOGY

The starting point of the CAMD methodology developed in this work is the methodology proposed by Serrato in 2009 [4]. This section is intended to compare the results obtained by the two methodologies and remark the improvements introduced in this work.

First of all, it must be pointed out that the results of the two methodologies are not comparable as the optimization problem formulated in the methodology of Serrato is intended to design solvents with optimal solvent power and the optimization problem in this work is intended to design solvents with optimal solvent power, environment-friendly and available in the market. In other words, while the methodology of Serrato addressed a chemical product design problem of one desirable features, the methodology of this work addressed a chemical product design of three desirable features.

The comparison done is based on the best designed solvents for the separation of lactic acid from an aqueous solution. For the solution of this problem Serrato reported a list of the 52 best solvents designed by executing 60 times the CAMD methodology proposed in his work. Table 29 present the original data reported and the SMILES representation of each solvent to make easier to identify the chemical structure. His reported solvents does not contain information about the solvent loss, then Table 30 contains the solvent loss of those solvents computed using the methods used in this work. This table also reports computed values of environmental-related properties, Environmental Index, the number of violated constraints and market availability for the solvents.

Table 29. Original set of optimal solvents, Serrato [4].

#	SMILES	Boiling Point	Melting point	Gibbs Energy	Density	KS
1	CC(C=O)(C=O)C=O	462.8	290.5	-309.8	1.21	1831.5
2	CC(CC=O)(C=O)C=O	481.5	295.9	-301.58	1.16	1622.76
3	C(C=O)C=O	386.5	230.7	-222.31	1.03	1658.81
4	CC(CCC=O)(C=O)C(=O)O	543.6	321.4	-532.78	1.24	578.37
5	CC(=O)OC(C)(C=O)C=O	474	292.1	-526.78	1.2	766.8
6	CC(=O)OCC(C)(C=O)C=O	469.5	287.8	-299.49	1.09	796.2
7	CC(C)(CCC(C)(CC=O)C=O)C=O	545.9	318.4	-263.204	1.04	558.12
8	CC(=O)OC(C)(CC=O)C=O	496.1	294.7	-512.341	1.16	791.5
9	CCCOC(=O)CC(C)(CCCC=O)C=O	562.4	313.7	-460.02	1.05	471.58
10	CC(=O)OCC(C)(CC(C)(C)C=O)C=O	540.5	315.4	-488.4	1.07	469.04
11	CC(=O)OC(C)(CCC(C)(C)C=O)CC=O	554.96	309.9	-478.811	1.06	458.22
12	CC(C=O)C=O	400.3	249.8	-220.085	1	833.5
13	CC(=O)OC(C)(CC(=O)O)C=O	541.6	316.2	-751.773	1.27	359.2
14	CC(CCC=O)(C=O)OC	487.7	293.3	-291.26	1.06	790.6
15	CC(C=O)(C=O)C(=O)O	515.4	312.9	-549.25	1.36	534.33
16	CC(=O)OC(C)(CCC=O)C(=O)O	554.5	320.5	-743.542	1.23	389.3
17	CCCOCC(C)(CCCC=O)OC(=O)C	549.1	305.3	-481.1	1.01	384.5
18	CCCCC(C)(CCOCC=O)OC(=O)C	550.6	294	-479.8	1.01	384.54
19	CC(CC=O)C=O	425.3	257.7	-211.854	0.98	697.3
20	CC(C)(CCC(C)(CC=O)C=O)OC	537.3	311.96	-261.113	1.01	437.1
21	CCCOC(=O)CC(C)(CCCC=O)OC(=O)C	572.3	312.7	-670.77	1.06	366.4
22	CCCNC(=O)OC(=O)C	546.4	338.7	-418.2	1.43	286.8
23	CC(=O)OCC(C)(CCC(C)(C)OC)C=O	548.5	310.9	-471.871	1.03	343.4
24	C=CCCC(C=O)(C(C)CC=O)OC	548.8	317.9	-182.45	1.09	334.6
25	COC(C=O)C=O	436.2	269.7	-317.8	1.16	759.93
26	CC(C)(CC(C)(CC=O)C=O)C=O	530.8	323.5	-272.8	1.06	584.5
27	CC(=O)OC(C)(C)CCCC(=O)OC=O	532.7	289.2	-700.319	1.11	307.83
28	CCCOCC(C)(CCCOC(=O)C)OC(=O)C	558	295.6	-696.76	1.02	304.3
29	CO(C)(C)(C=O)COC(=O)CC(=O)O	564.5	310.5	-722.785	1.19	304.8
30	CC(C)(CCC(C)(CC(=O)O)C=O)OC	577.1	342.4	-502.694	1.08	295.6
31	CCCCC(=O)OC(C)(CCCOC(=O)C)OC(=O)C	580.3	303.7	-886.38	1.07	293.3
32	CCCCC(C)(C)OCCOC(=O)CC=O	545.2	300	-469.98	1	290.3
33	CC(=O)NCCOC(=O)C	546.4	174.1	-418.2	1.43	286.8
34	CCC(C)(CCC(=O)OC(C)(C)OC(=O)C)C=O	563.4	311.8	-673.553	1.07	285.2
35	CCCCC(C)(COCCOC)OC(=O)C	542.1	285.7	-477.7	0.98	313.43
36	COCC=O	367.1	214.8	-220.222	0.95	517.8
37	CC(C)(C=O)C=O	416.2	264.4	-209.99	0.99	516.1
38	CC(C)(CCC(C)(C=O)CO)OC	541.9	301.9	-287.502	1.03	138.4
39	CC(C)(CC(C)(CC(=O)O)C=O)C=O	571.2	332.3	-510.867	1.14	347
40	CC(C=O)(C=O)O	470.3	286.2	-360.5	1.5	386.6
41	CC(C=O)(C=O)OC	449.7	281.9	-307.72	1.13	774.1
42	CC(C)(CCC(C)(C=O)CO)C=O	542.6	308.5	-311.844	1.09	195.27
43	CC(=O)OC(C)(C)CNC(=O)CCC(=O)O	634.2	290	-727.286	1.45	265.98
44	CC(C=O)(OC)OC	435.6	272.5	-305.6	1.06	470.1
45	CC(=O)NCCCCOC(=O)C	570.9	203.9	-401.7	1.29	328.08
46	CCCOCC(C)(CCCO)OC(=O)C	555.5	289.9	-530.52	1.08	262.36
47	CC(=O)OC(C)(CC(=O)O)CC(C)(C)C=O	579.2	324.9	-726.47	1.15	310.9
48	CC(C=O)(C=O)O	470.3	286.2	-360.54	1.5	386.6
49	CCCC(=O)NCOC(=O)C	559	190	-409.94	1.35	314.3
50	CC(=O)OC(C)(CC(C)(C)C=O)CO	547.2	301.5	-537.773	1.17	263.7
51	CCC(C)(CCOC(C)(C)C=O)C(=O)O	568.3	315.8	-511.2	1.06	240.02
52	CCC(C)(CC(=O)NCC(C)(C)OC(=O)C)OC(=O)C	632.2	281.7	-682.9	1.24	285.5

Table 30. Extra properties computed for Serrato optimal solvents.

#	LC50FM (log)	LC50DM (log)	LD50 (log)	Water Solubilty (log)	BFC (log)	Solvent Loss	Environ-mental Index	Penal-ties	Market Available
1	-4.14	-5.66	-1.55	-0.50	-1.32	0.24	13.64	1	no
2	-4.29	-5.83	-1.55	-0.52	-1.22	0.16	14.17	2	no
3	-3.69	-4.65	-1.58	0.48	-0.62	0.56	15.48	1	yes
4	-3.07	-4.85	-1.81	-0.21	-1.91	0.12	11.82	2	no
5	-3.80	-4.24	-1.56	-0.40	-0.56	0.09	14.14	0	no
6	-4.13	-4.41	-1.46	-0.91	-0.47	0.05	13.74	0	no
7	-4.68	-6.56	-1.60	-2.06	-0.61	0.00	15.27	1	no
8	-3.95	-4.41	-1.55	-0.41	-0.47	0.05	14.69	1	no
9	-4.98	-6.12	-1.50	-2.20	0.42	0.00	17.30	1	no
10	-4.23	-4.97	-1.51	-2.10	0.05	0.00	14.55	1	no
11	-4.34	-5.14	-1.60	-1.91	0.14	0.00	15.82	1	no
12	-4.02	-4.27	-1.60	-0.52	-0.44	0.33	14.78	1	yes
13	-3.24	-3.26	-1.69	-0.47	-0.99	0.09	12.02	1	no
14	-3.31	-6.04	-1.65	-0.66	-0.52	0.06	15.72	1	no
15	-2.64	-4.51	-1.81	0.13	-2.10	0.45	10.97	2	no
16	-2.73	-3.43	-1.81	-0.08	-1.16	0.05	12.35	1	no
17	-4.24	-4.54	-1.68	-1.77	0.83	0.00	17.63	1	no
18	-4.24	-4.54	-1.68	-1.77	0.83	0.00	17.63	1	no
19	-4.17	-4.44	-1.59	-0.56	-0.34	0.16	15.30	1	yes
20	-3.41	-6.60	-1.70	-1.86	0.00	0.00	16.52	1	no
21	-4.64	-4.71	-1.50	-2.04	1.17	0.00	17.86	1	no
22	-3.92	-6.19	-1.62	-0.68	-0.61	1.29	16.06	2	no
23	-3.25	-5.18	-1.61	-2.22	0.75	0.00	16.13	1	no
24	-4.65	-6.37	-1.81	-1.95	-0.02	0.01	17.93	1	no
25	-3.41	-5.11	-1.62	-0.05	-0.85	0.63	14.57	1	yes
26	-4.39	-6.39	-1.60	-1.74	-0.71	0.00	14.96	1	no
27	-3.77	-4.19	-1.50	-1.11	0.89	0.00	16.89	0	no
28	-4.08	-3.13	-1.59	-2.12	1.58	0.00	17.24	1	no
29	-2.83	-4.61	-1.32	0.24	0.68	0.04	16.52	1	no
30	-2.70	-5.44	-1.84	-1.91	-0.53	0.00	13.87	2	no
31	-4.48	-3.29	-1.42	-2.39	1.93	0.00	17.48	2	no
32	-3.83	-5.39	-1.63	-1.97	1.24	0.00	18.66	1	no
33	-3.92	-6.19	-1.62	-0.68	-0.61	1.29	16.06	1	yes
34	-3.73	-4.58	-1.56	-2.28	1.31	0.00	17.17	1	no
35	-3.12	-4.58	-1.79	-1.88	1.44	0.00	18.64	0	no
36	-2.56	-4.69	-1.68	0.35	-0.01	0.57	16.48	1	yes
37	-3.50	-4.86	-1.63	-0.79	-0.29	0.14	15.08	1	yes
38	-2.20	-4.96	-1.66	-1.75	0.32	0.00	14.48	1	no
39	-3.68	-5.23	-1.74	-1.80	-1.24	0.01	12.30	1	no
40	-2.85	-4.02	-1.60	0.66	-1.17	0.38	12.83	1	no
41	-2.88	-5.70	-1.65	-0.31	-0.71	0.33	14.88	1	no
42	-3.46	-4.92	-1.55	-1.95	-0.29	0.00	13.23	1	no
43	-3.78	-6.39	-1.83	-1.12	-1.64	0.09	13.75	1	no
44	-1.61	-5.74	-1.76	-0.11	-0.10	0.25	16.13	1	yes
45	-4.50	-6.54	-1.62	-1.33	-0.42	0.28	16.68	1	yes
46	-3.03	-2.91	-1.64	-1.67	1.15	0.00	15.58	0	no
47	-3.34	-3.82	-1.74	-1.65	-0.48	0.01	12.85	2	no
48	-2.85	-4.02	-1.60	0.66	-1.17	0.38	12.83	1	no
49	-4.21	-6.37	-1.62	-1.01	-0.51	0.57	16.37	1	no
50	-2.84	-3.34	-1.56	-1.49	0.37	0.01	13.46	1	no
51	-2.16	-4.68	-1.99	-1.49	-0.58	0.00	13.75	1	no
52	-4.11	-6.52	-1.74	-2.08	0.05	0.00	17.26	1	no

The most remarkable feature of Table 30 is that only five compounds meet the original constraints of the problem and none of them are available in the market. The five compounds are highlighted in the table.

For the purpose of inspecting the effect of including environmental issues and market availability in the design of optimal solvents by CAMD, Figure 47 shows the distribution of the Yield KS for the solvents designed by Serrato and the solvents designed by the methodology of this work (those shown in Figure 42). The average of the solvents of this work is lower compared to the average of the solvents of Serrato, therefore the inclusion of environment-related objectives and market availability in the optimization has a negative impact over the solvent power of the designed solvents.

The standard deviation of the Yield KS of the designed solvents in this work is larger than the standard deviation of the solvents designed by Serrato. This causes that Yield KS looks more spread for the solvents designed in this work. This behaviour is a consequence of optimizing environment-related issues at the same time Yield KS is being optimized.

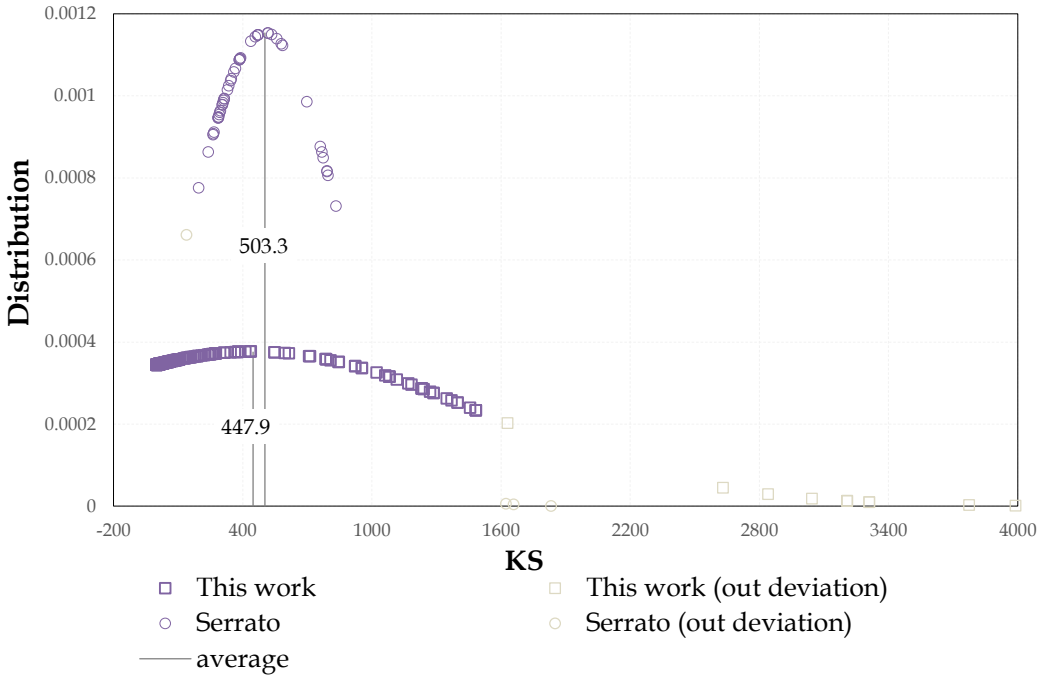


Figure 47. Distribution of Yield for Serrato and this work best solvents.

Regarding the environmental performance of the solvents designed by both methodologies, Figure 48 shows the distribution of the Environmental Index computed for the solvents designed by Serrato and the distribution of the solvents designed by the methodology of this work (those shown in Figure 42). As expected, the average of the solvents of this work is lower compared to the average of the solvents designed by Serrato. In addition, the standard deviation of the solvents of this work is larger than the standard deviation of the solvents of Serrato. As a consequence, the Environmental Index is more spread for the solvents designed in this work and the selection of solvents in the bounds with good environment-friendly features is possible.

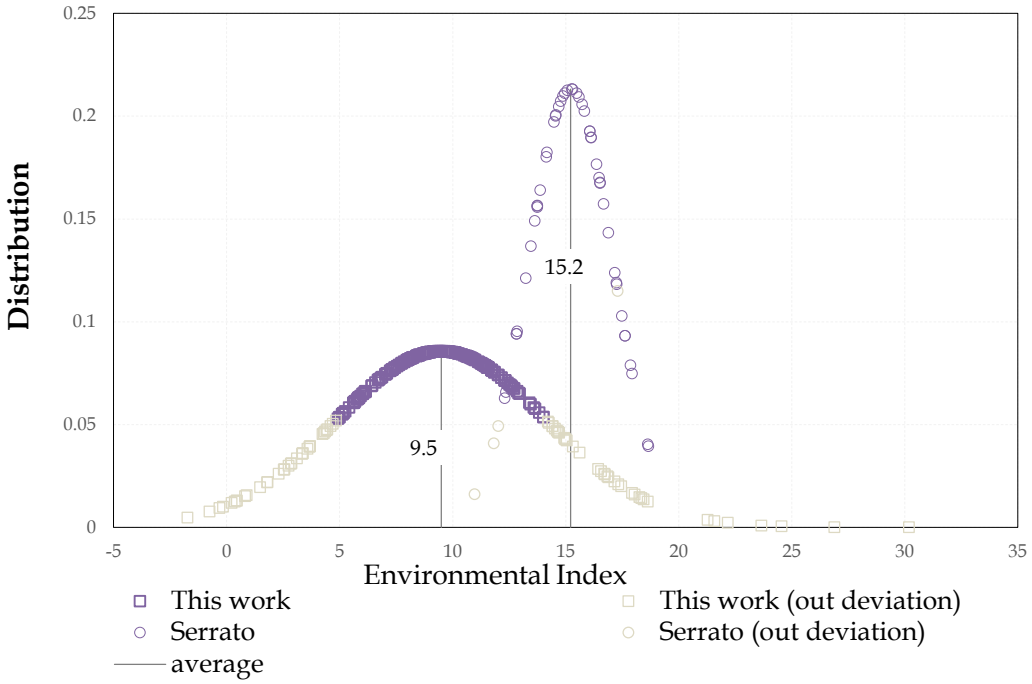


Figure 48. Distribution of Environment index for Serrato and this work best solvents.

5 Conclusions and future work

5.1 Conclusions

In this work, the design of liquid-liquid extraction solvents with environment-friendly features and availability in the market is successfully achieved. A novel CAMD methodology based on multi-objective optimization is proposed and implemented. The study case is the separation of lactic acid from an aqueous solution.

The environment-related objective is easy to reach in the presented CAMD methodology. In contrast, there are no optimal values for the solvent-power objective and the continuous growth of this objective is only limited by the constraints of the optimization problem.

Accuracy of methods for estimating environment-related properties can lead to reduce the confidence of environmental objectives in a CAMD methodology. Prediction of environment-related properties is still a matter that requires development of more rigorous methods, as most of the available methods do not count with very good accuracy. The reported values of absolute error for the environment-related estimation methods used in this work lie near 0.4-0.5 in logarithmic scale, which is significant.

Chemical product design of compounds with multiple desirable properties including mixture properties is difficult. Handling these properties can be challenging as those depend not only on the candidate compound but also on its interactions with the medium. In the case of solvent loss, a greater weight for the constraint related to this property is necessary to reach an acceptable number of feasible solvents designed.

Market availability favours the design of environment-friendly solvents. In addition, there is a trade-off between market availability and solvent-power.

Regarding the design of solvents for the separation of lactic acid from an aqueous solution, the task of the decision maker is difficult. The executions of the CAMD program produces a concave Pareto best front for this problem, which is an unfavourable result as it reduces the number of solutions with a good balance among the optimization objectives.

Designing environment-friendly solvents without significantly reducing the separation power is possible. Compared to the solvents designed by the CAMD methodology of Serrato, the solvents designed by the methodology of this work present a notorious improvement for the environmental performance. In the case

of solvent-power, the solvents designed by Serrato present better values for this property, however the difference is not great.

Molecules are sensitive systems and modifications involving more than one functional group can lead to radical changes in the properties. The CAMD methodology of this work includes six genetic operators to evolve molecules and the best performance was shown by the less strong operators. Namely, the chain extension operator and the group removal operator presents the mentioned best performance. On the other hand, very strong genetic operators present poor performance. In the case of the Chain replacement operator, the strongest genetic operator, it presented the worst performance.

To conduct the process of designing optimal solvents, the Multi-Objective Hybrid Adaptive Evolutionary Algorithm (MOHAEA) is proposed and implemented in this work. In this algorithm, genetic operators count with variable application rates as well as the original HAEA. The fitness of each individual is compute by combining Pareto dominance ranks and crowding-distance. The constraints handling is done by a list of weights where the punishment for an unfeasible solution consists in reducing the Pareto dominance of the solution.

The usage of Pareto dominance in the evaluation of the solutions reduces the bias of the optimization. Furthermore, the proposed weighting system for the constraints in MOHAEA eases the handling of constraints, as each weight does not depend on the numerical nature of the constraint but on the importance of each constraint. If all constraints are equally important, the default value of 1.0 is a good choice.

A comprehensive evaluation of MOHAEA is not possible in this work, as it requires testing the algorithm on many optimization problems and that is out of the scope. Moreover, the proposed CAMD methodology addresses a combinatorial problem of not known solution where one of the objectives does not have a convergence range. No general conclusions can result from a problem with those features.

5.2 FUTURE WORK

The CAMD methodology proposed in this work produces satisfactory results for the design of environment-friendly separation solvents. However, further opportunities for improving the quality of the results and increasing the capabilities of the program are listed in this section. The most important recommendations for further work are presented as follows:

- The inclusion of aromatic and cyclic functional groups. The new methodology presented as well as the methodology of Serrato excludes those groups from the molecule search space, although many compounds containing these groups are of importance in industry.

- A more rigorous study of the constraints. Currently the majority of the designed solvents are not feasible and the optimization of one objective function depends on the constraints. A parametric analysis of the constraint weights can be worthwhile for the reduction of unfeasible solutions.

- Experimental validation of the results. Solvents with market availability were designed in this work, then the experimental validation is possible. Laboratory work was never included in the scope of this work.

- A comprehensive evaluation of the MOHAEA algorithm. In this work, the original HAEA algorithm was modified in order to address multi-objective optimization problems and a complete study of the capabilities of the new algorithm has not been performed.

- Testing the CAMD methodology in more solvents design problems. This work only used as study case the separation of lactic acid from an aqueous solution, which is a valid case, but it would be interesting to check the performance of the new methodology in more separation problems.

- Evaluation of new objective functions. One of the objective functions in the proposed CAMD methodology does not count with a convergence point, making very difficult the definition of optimal solutions and the study of the optimization method, hence the evaluation of new objective functions able to represent the solvent power may lead to define a friendlier optimization problem.

Nomenclature

s	Separation agent, solvent to design
t	Target solute, compound to separate
p	Problem solvent containing the target solute
$P_{i,\alpha/\beta}$	Partition coefficient of the compound i in the phases α and β
$\gamma_{i,j}^{\infty}$	Activity coefficient of i the phase j at infinite dilution
MW_i	Molecular weight of the compound i
$x_{i,j}$	Mole fraction of i dissolved in the phase j
K	Extraction power
S	Selectivity
ΔG_f	Standard Gibbs energy of formation
T_b	Normal boiling point
T_m	Normal melting point
V_m	Liquid molar volume
ρ	Density
T	Temperature
LC_{50}	Lethal Concentration
LC_{50}^{FM}	*Fathead Minnow* 96hr lethal concentration
LC_{50}^{DM}	*Daphnia Magna* 48hr lethal concentration
LD_{50}	Oral rat lethal dose
W_s	Water solubility
BCF	Bioconcentration factor
k_{LD50}	Environmental index constant for LD_{50}
k_{WS}	Environmental index constant for WS
k_{BCF}	Environmental index constant for BCF
A	Market availability category
Y	Solvent yield
E	Environmental index
N_i	Number of occurrences of first-order groups
N_j	Number of occurrences of second-order groups
C_i	Contribution of first-order groups to a property.
C_j	Contribution of second-order groups to a property.
ARE	Average Relative Error
AAE	Average Absolute Error
SD	Standard deviation
$i_{crowd-dist}$	Crowding distance measure for the individual i
$i_{distance}$	Diversity distance component for the individual i
i_{rank}	Pareto rank of the individual i

References

[1] R. G. a. V. V. L.E.K. Achenie, «Chapter 1. Introduction to CAMD,» de *Computer Aided Molecular Design: Theory and Practice*, Elsevier Science B.V, 2003, pp. 3-21.

[2] N. D. Austin y N. V. Sahinidiscorrespondence, «Computer-aided molecular design: An introduction and review of tools, applications, and solution techniques,» *Chemical Engineering Research and Design,* 2016.

[3] G. D. M. E. L. Cussler, Chemical Product Design, Cambridge, UK: Cambridge University Press, 2011.

[4] J. C. Serrato Bermúdez, «Diseño Computacional de Agentes de extracción para la Separación de Compuestos Orgánicos en Corrientes Acuosas- Aplicación al Ácido Láctico,» Universidad Nacional de Colombia, Bogotá, 2009.

[5] L. A. Cisternas, Gálvez y E. D., « Principles for chemical products design,» *Computer Aided Chemical Engineering,* vol. 21, pp. 1107-1113, 2006.

[6] G. D. M. E. L. Cussler, «An introduction to chemical product design,» *Trans IChemE,* vol. 78, pp. 5-11, 2000.

[7] L. Ng, F. Chong y N. Chemmangattuvalappil, «Challenges and opportunities in computer-aided molecular design,» *Computers and Chemical Engineering,* vol. 81, pp. 115-129, 2015.

[8] M. Edena, S. Jørgensena y R. Gani, «A novel framework for simultaneous separation process and product design,» *Chemical Engineering and Processing: Process Intensification,* vol. 43, nº 5, pp. 595-608, 2004.

[9] O. Odele y S. Macchietto, «Computer aided molecular design: a novel method for optimal solvent selection,» *Fluid Phase Equilibria,* vol. 82, pp. 47-54, 1993.

[10] L. Ng y N. Chemmangattuvalappil, «Tools for Chemical Product Design. Chapter 1 – Mathematical Principles of Chemical Product Design and Strategies,» de *Tools for Chemical Product Design*, Auburn, Alabama, United States, Elsevier, 2017, pp. 3-44.

[11] R. Gani, B. Nielsen y A. Fredenslund, «A group contribution approach to computer-aided molecular design,» *AIChE Journal,* vol. 37, nº 9, pp. 1318-1332, 1991.

[12] R. Gani, «Chemical product design: challenges and opportunities,» *Computers and Chemical Engineering,* vol. 28, pp. 2441-2457, 2004.

[13] H. P.M y R. Gani, « A multi-step and multi-level approach for computer aided molecular design,» *Computers and Chemical Engineering,* vol. 24, pp. 677-683, 2000.

[14] B. C. Roughton, «Development of Computer-Aided Molecular Design Methods for Bioengineering Applications,» University of Kansas, Kansas, United States, 2013.

[15] S. Macchietto, O. Odele y O. Omatsone, «Design on optimal solvents for liquid-liquid extraction and gas absorption processes,» *Chemical engineering research & design* , vol. 68, pp. 429-433, 1991.

[16] P. M. Harper, R. Gani y P. Kolar, «Computer-aided molecular design with combined molecular modeling and group contribution,» *Fluid Phase Equilibria,* vol. 158, pp. 337-347, 1999.

[17] K. Joback y R. Reid, «Estimation of pure-component properties from group-contribution,» *Chemical Engineering Communications,* vol. 57, pp. 233-243, 1987.

[18] L. G. R. Constantinou, « New group contribution method for estimating properties of pure compounds,» *AIChE Journal,* vol. 40, pp. 1697-1710, 1994.

[19] J. Marrero y R. Gani, «Group-contribution based estimation of pure component properties,» *Group-contribution based estimation of pure component properties,* vol. 183, pp. 183-208, 2001.

[20] D. Bonchev, Chemical graph theory: Introduction and fundamentals, Paris, France: CRC Press, 1991.

[21] H. Wiener, «Structural determination of paraffin boiling points,» *Journal of the American Chemical Society,* vol. 1, pp. 17-20, 1947.

[22] M. Randic, Z. Mihalic y N. S. , «Graphical bond orders: novel structural descriptors,» *Journal of Chemical Information and Computer Sciences,* vol. 34, pp. 403-409., 1994.

[23] L. Kier, «A shape index from molecular graphs,» *Quantitative Structure-Activity Relationships,* vol. 4, pp. 109-117, 1985.

[24] K. Camarda y C. Maranas, «Optimization in polymer design using connectivity indices,» *Industrial and Engineering Chemistry Research,* vol. 38, pp. 1884-1892, 1999.

[25] D. Visco, R. Pophale y J.-L. Faulon, «Developing a methodology for an inverse quantitative structure-activity relationship using the signature molecular descriptor.,» *Journal of Molecular Graphics and Modelling,* vol. 20, pp. 429-438, 2002.

[26] C. D. Maranas, «Optimal computer-aided molecular design: A polymer design case study,» *Industrial & Engineering Chemistry Research,* vol. 35, pp. 3403-3414, 1996.

[27] S. Burer y A. Letchford, «Non-convex mixed-integer nonlinear programming: a survey,» *Surveys in Operations Research and Management Science ,* vol. 17, pp. 97-106, 2012.

[28] F. Glover, «Future paths for integer programming and links to artificial intelligence,» *Computers and Operations Research,* vol. 13, pp. 533-549, 1986.

[29] M. Melanie, «Chapter 1: Genetic Algorithms: An Overview,» de *An Introduction to Genetic Algorithms*, vol. 7, London, England, MIT Press, 1996, pp. 2-26.

[30] V. Venkatasubramanian, K. Chan y C. J. M., «Evolutionary design of molecules with desired properties using the genetic algorithm.,» *Journal of Chemical Information and Computer Sciences,* vol. 35, pp. 188-195, 1995.

[31] B. Van Dyk y I. Nieuwoudt, «Design of solvents for extractive distillation,» *Industrial & Engineering Chemistry Research,* vol. 39, p. 1423–1429, 2000.

[32] J. E. Ourique y A. S. Telles, «Computer-aided molecular design with simulated annealing and molecular graphs,» *Computers & Chemical Engineering,* vol. 22, pp. 56-15-5618, 1998.

[33] B. Lin, S. Chavali y K. Camarda, «Computer-aided molecular design using Tabu search,» *Computers & Chemical Engineering,* vol. 29, pp. 337-347, 2005.

[34] R. Gani y A. Fredenslund, «Computer-aided molecular and mixture design with specified constraints,» *Fluid Phase Equilibria,* vol. 82, pp. 39-46, 1993.

[35] M. R. Eden, S. B. Jørgensen y R. Gani, «A novel framework for simultaneous separation process and product design,» *Chemical Engineering and Processing: Process Intensification,* vol. 43, pp. 595-608, 2004.

[36] D. A.P. y A. L.E., «Designing environmentally safe refrigerants using mathematical programming,» *Chemical Engineering Science,* vol. 15, pp. 3727-3739, 1996.

[37] E. Pistikopoulos y S. Stefanis, «Optimal solvent design for environmental impact minimization,» *Computational Chemical Engineering,* pp. 717-733, 1998.

[38] J. Ten, M. Hassim y D. Ng, «A molecular design methodology by the simultaneous optimisation of performance, safety and health aspects,» *Chemical Engineering Science,* vol. 159, pp. 140-153, 2017.

[39] A. Hukkerikar, S. Kalakul y R. Gani, «Estimation of environment-related properties of chemicals for design of sustainable processes: development of group-contribution+ (GC+) property models and uncertainty analysis.,» *Chemical Information Modelling,* vol. 11, pp. 2823-39, 2012.

[40] R. T. Ng, M. H. Hassim y D. K. Ng, «Process synthesis and optimization of a sustainable integrated biorefinery via fuzzy optimization.,» *AIChE Journal,* vol. 59, pp. 4212-4227, 2013.

[41] J. Ooi, M. Angelo B. Promentilla y R. Tan, «Alternative Solvent Design for Oil Extraction from Palm Pressed Fibre via Computer-Aided Molecular Design,» de *Green Technologies for the Oil Palm Industry*, Singapore, Springer Singapore, 2019, pp. 33-55.

[42] P. C. Fishburn, « Additive utilities with incomplete product sets: Application to priorities and assignments,» *Operations Research,* vol. 15, pp. 537-542, 1967.

[43] M. G. X. 2. Ehrgott, «Multiple Criteria Optimization,» de *Multiple Criteria Optimization: State of the Art Annotated Bibliographic Surveys.* , USA, Kluwer Academic Publishers, 2002.

[44] R. Marler y J. Arora, «Survey of multi-objective optimization methods for engineering,» *Structural and Multidisciplinary Optimization,* vol. 26, nº 6, pp. 369-395, 2004.

[45] H. Rice, «Appendix 12b. Nomenclature | Liquid-Liquid Distribution (Solvent Extraction),» de *Encyclopedia of Separation Science*, Cheshire, UK, Academic Press, 2000, pp. 4753-4772.

[46] S. James, «Octanol-Water Partition Coefficients of Simple Organic Compounds,» *J. Phys. Chern. Ref. Data,,* pp. 1111-1227, 1989.

[47] E. Brignole y S. Pereda , Phase Equilibrium Engineering, Bahía Blanca, Argentina: Elsevier, 2013.

[48] D. Johnson y G. Stracher, Thermodynamic Loop Applications in Materials Systems, Minerals, Metals & Materials Society, 1995, pp. 61-79.

[49] EPA, «Health Effects Test Guidelines OPPTS 870.1300 Acute Inhalation Toxicity,» 1998.

[50] EPA, «Health Effects Test Guidelines OPPTS 870.1100 Acute Oral Toxicity,» 2002.

[51] EPA, «Ecological Effects Test Guidelines OCSPP850.1730: Fish Bioconcentration Factor (BCF),» 2016.

[52] B. K. Shoichet and J. J. Irwin, "ZINC − A Free Database of Commercially Available Compounds for Virtual Screening," *Journal of Chemical Information and Modeling,* pp. 177-182, 2005.

[53] UCSF, «ZINC,» 11 May 2019. [En línea]. Available: http://zinc15.docking.org.

[54] ZINC, «Tranches - ZINC,» 12 May 2019. [En línea]. Available: http://zinc15.docking.org/tranches/home/.

[55] ZINC, «ZINC,» 30 March 2018. [En línea]. Available: http://zinc15.docking.org/tranches/home/.

[56] EPA, «Distributed Structure-Searchable Toxicity (DSSTox) Database,» 12 May 2019. [En línea]. Available: https://www.epa.gov/chemical-research/distributed-structure-searchable-toxicity-dsstox-database.

[57] A. J. Williams y C. M. Grulke, «The CompTox Chemistry Dashboard: a community data resource for environmental chemistry,» *Journal of Cheminformatics,* p. 9:61, 2017.

[58] EPA, «CompTox,» 12 April 2018. [En línea]. Available: https://comptox.epa.gov/dashboard/downloads.

[59] A. D. McEachran y J. R. Sobus, «Identifying known unknowns using the US EPA's CompTox Chemistry Dashboard,» *Analytical and Bioanalytical Chemistry,* pp. 1729-1735, 2017.

[60] A. S. Hukkerikar, B. Sarup, A. Ten y R. Gani, «Group-contribution (GC +) based estimation of properties of pure components: Improved property estimation and uncertainty analysis,» *Fluid Phase Equilibria,* vol. 321, pp. 25-43, 2012.

[61] A. S. Hukkerikar, G. Sin, J. Abildskov, B. Sarup and R. Gani, Development of pure component property models for chemical product-process design and analysis, Kgs. Lyngby, Denmark: Technical University of Denmark, Department of Chemical and Biochemical Engineering, 2013.

[62] tutorialspoint.com, «MongoDB - Overview,» [En línea]. Available: https://www.tutorialspoint.com/mongodb/mongodb_overview.htm. [Último acceso: 02 June 2019].

[63] D. Weininger, «SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules,» *Journal of Chemical Information and Computer Sciences,* vol. 28, nº 1, pp. 31-36, 1988.

[64] Daylight Chemical Information Systems, «3. SMILES - A Simplified Chemical Language,» [En línea]. Available: https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html. [Último acceso: 02 June 2019].

[65] N. M. O'Boyle, «Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI,» *Journal of cheminformatics,* vol. 4, nº 1, p. 22, 2012.

[66] «Tranche Browser,» [En línea]. Available: http://zinc15.docking.org/tranches/home/. [Último acceso: 19 March 2018].

[67] T. Sterling y J. J. Irwin, «ZINC 15 - Ligand Discovery for Everyone,» *Journal of chemical information and modeling,* vol. 55, nº 11, pp. 2324-2337, 2015.

[68] E. L. Willighagen, J. W. Mayfield, J. Alvarsson y A. Berg, «The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching,» *Journal of Cheminformatics,* vol. 9, nº 1, p. 33, 2017.

[69] InChI Trust, «InChI and InChIKeys for chemical structures,» [En línea]. Available: https://www.inchi-trust.org/. [Último acceso: 12 July 2019].

[70] D. Constantinescu y J. Gmehling, «Further Development of Modified UNIFAC (Dortmund): Revision and Extension 6,» *Journal of Chemical & Engineering Data,* vol. 61, nº 8, pp. 2738-612748, 2016.

[71] V. Papaioannou, G. Jackson, G. Jackson y A. Galindo, «Group Contribution Methodologies for the Prediction of Thermodynamic Properties and Phase Behavior in Mixtures,» de *Process Systems Engineering: Vol. 6 Molecular Systems Engineering*, vol. 6, Weinheim, Germany, Wiley Online Books, 2011, pp. 135-172.

[72] J. Smith, H. Van Ness y M. Abbott, «Apéndice H. Método Unifac,» de *Introducción a la termondinámica en ingeniería química*, México, D.F., México, McGraw-Hill, 2005, pp. 791-797.

[73] J. Gmehling, D. Constantinescu y B. Schmid, «Group Contribution Methods for Phase Equilibrium Calculations,» *Annual Review of Chemical and Biomolecular Engineering,* vol. 6, nº 1, pp. 267-292, 2015.

[74] DDBST GmbH, «Parameters of the Modified UNIFAC (Dortmund) Model,» [En línea]. Available: unifac.ddbst.de/PublishedParametersUNIFACDO.html. [Último acceso: 20 05 2019].

[75] K. Deb, L. Wang y A. H. C. Ng, «Multi-objective Optimisation Using Evolutionary Algorithms: An Introduction,» de *Multi-objective Evolutionary Optimisation for Product Design and Manufacturing*, London, UK, Springer London, 2011, pp. 3-34.

[76] J. Gomez, «Self Adaptation of Operator Rates in Evolutionary Algorithms,» *Genetic and Evolutionary Computation – GECCO 2004,* pp. 1162-1173, 2004.

[77] N. Riquelme, C. Von Lücken y B. Baran, «Performance metrics in multi-objective optimization,» *2015 Latin American Computing Conference (CLEI),* pp. 1-11, 2015.

[78] K. Deb, A. Pratap, S. Agarwal y T. Meyarivan, «A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II,» *IEEE Transactions on Evolutionary Computation,* vol. 6, nº 2, pp. 182-197, 2002.

[79] S. Rodrigues, P. Bauer y P. A. Bosman, «Multi-objective optimization of wind farm layouts – Complexity, constraint handling and scalability,» *Renewable and Sustainable Energy Reviews,* vol. 65, pp. 587-609, 2016.

[80] N. Trinajstic, Chemical Graph Theory, Zagreb, Croatia: CRC Pres, 1992.

[81] DDBST GmbH, «Parameters of the Modified UNIFAC (Dortmund) Model,» [En línea]. Available: http://unifac.ddbst.de/PublishedParametersUNIFACDO.html#ListOfSubGroupsAndTheirGroupSurfacesAndVolumes. [Último acceso: 01 June 2019].

[82] DDBST GmbH , «UNIFAC Consortium.,» [En línea]. Available: http://unifac.ddbst.de/unifac_.html. [Último acceso: 01 June 2019].

[83] K. A. Rodríguez R., 2019.

[84] United States Environmental Protection Agency , «Toxicity Estimation Software Tool (TEST),» [En línea]. Available: https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test. [Último acceso: 23 June 2019].

[85] W.-H. Lee, «EPI Suite™-Estimation Program Interface,» United States Environmental Protection Agency, 2012. [En línea]. Available: https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface. [Último acceso: 21 June 2019].

[86] R. H. Perry y D. W. Green, «Section 2. Physical and Chemical Data,» de *Perry's Chemical Engineers' Handbook*, New York, McGraw-Hill, 2008, pp. 195-200.

[87] Y. Zhang, H. Wang, Z. Yang y J. Li, «Relative Accuracy Evaluation,» *PLoS One,* vol. 9, nº 8, p. e103853, 2014.

[88] T. Okuyama y M. Maskill, «3.9. Intermolecular Interactions and Physical Properties of Organic Compounds,» de *Organic Chemistry : A Mechanistic Approach*, Oxford, UK, Oxford University Press, 2013, pp. 61-66.

[89] S. Terasawa, H. Itsuki y S. Arakawa, «Contribution of hydrogen bonds to the partial molar volumes of nonionic solutes in water,» *The Journal of Physical Chemistry,* vol. 79, nº 22, pp. 2345-2351, 1975.

[90] United States Environmental Protection Agency, «ECOTOX Knowledgebase,» 23 June 2019. [En línea]. Available: https://cfpub.epa.gov/ecotox/search.cfm.

[91] U.S. National Institutes of Health, «ChemIDplus Advanced. A TOXNET database,» [En línea]. Available: https://chem.nlm.nih.gov/chemidplus/chemidheavy.jsp.

[92] J. A. Arnot y F. A. Gobas, «A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms,» *Environmental Reviews,* vol. 14, nº 4, pp. 257-297, 2006.

[93] Ambit Project, «EURAS bioconcentration factor (BCF) Gold Standard Database,» 16 May 2018. [En línea]. Available: http://ambit.sourceforge.net/euras/. [Último acceso: 23 June 2019].

[94] U.S. EPA/National Risk Management Research, User's Guide for T.E.S.T. (version 4.2), Cincinnati, U.S.: U.S. Environmental Protection Agency , 2016.

[95] «Normalizing Data,» de *Encyclopedia of research design*, Australia, Thousand Oaks, Calif, 2010, pp. 1-4.

[96] B. E. Mitchell y P. C. Jurs, «Prediction of Infinite Dilution Activity Coefficients of Organic Compounds in Aqueous Solution from Molecular Structure,» *Journal of Chemical Information and Computer Sciences,* vol. 38, nº 2, pp. 200-209, 1998.

[97] J. Gmehling, M. Kleiber, . B. K. Kolbe y J. Rarey, «Phase Equilibria in Fluid System,» de *Chemical Thermodynamics for Process Simulation*, Fußgönheim, Germany, Wiley-VCH, 2019, pp. 173-322.

[98] S. Obayashi, D. Sasaki y A. Oyama, «Finding Tradeoffs by Using Multiobjective Optimization Algorithms,» *Transactions of The Japan Society for Aeronautical and Space Sciences - TRANS JPN SOC AERON SPACE SCI,* vol. 47, nº 155, pp. 51-58, 2004.