DE GRUYTER OPEN

HJHUNGARIAN JOURNAL OF INDUSTRY AND CHEMISTRY
Vol. 44(1) pp. 39–49 (2016)
hjic.mk.uni-pannon.hu
DOI: 10.1515/hjic-2016-0005

HUNGARIAN JOURNAL OF INDUSTRY AND CHEMISTRY
HJIC

# GROUP CONTRIBUTION METHOD-BASED MULTI-OBJECTIVE EVOLUTIONARY MOLECULAR DESIGN

GYULA DÖRGŐ[1] AND JÁNOS ABONYI* [1,2]

[1] Department of Process Engineering, University of Pannonia, Egyetem str. 10, Veszprém, H-8200, HUNGARY
[2] Institute of Advanced Studies Kőszeg, Chernel str. 14, Kőszeg, H-9730, HUNGARY

The search for compounds exhibiting desired physical and chemical properties is an essential, yet complex problem in the chemical, petrochemical, and pharmaceutical industries. During the formulation of this optimization-based design problem two tasks must be taken into consideration: the automated generation of feasible molecular structures and the estimation of macroscopic properties based on the resultant structures. For this structural characteristic-based property prediction task numerous methods are available. However, the inverse problem, the design of a chemical compound exhibiting a set of desired properties from a given set of fragments is not so well studied. Since in general design problems molecular structures exhibiting several and sometimes conflicting properties should be optimized, we proposed a methodology based on the modification of the multi-objective Non-dominated Sorting Genetic Algorithm-II (NSGA-II). The originally huge chemical search space is conveniently described by the Joback estimation method. The efficiency of the algorithm was enhanced by soft and hard structural constraints, which expedite the search for feasible molecules. These constraints are related to the number of available groups (fragments), the octet rule and the validity of the branches in the molecule. These constraints are also used to introduce a special genetic operator that improves the individuals of the populations to ensure the estimation of the properties is based on only reliable structures. The applicability of the proposed method is tested on several benchmark problems.

**Keywords:** computer-aided molecular design, multi-objective optimization, evolutionary algorithm, the Joback method, soft constraints

## 1. Introduction

The search for compounds exhibiting the desired physical and chemical properties is of significant industrial importance in the search for different chemicals and materials such as polymers [1, 2], blends [3], coatings, solvents, inert agents, heat transfer media [4], and drugs [5]. In the well-known technologies of the chemical, petrochemical, and pharmaceutical industries, the used medium for the given tasks has been developed via practical experience. For the improvement of these technologies or the design of a new process, every hypothetical molecule must be synthesized and tested to check the fulfillment of the design properties. This 'trial and error'-type method for the search of the appropriate agent with the defined properties is slow, inefficient and expensive, thus infeasible in the modern chemical industry and research. However, the problem is complex; the algorithmization can be carried out extensively with the use of property estimation methods and molecular structural feasibility operators. Algorithmic problems can be efficiently solved with the tools of process engineering. Over recent decades, the search for new compounds has resulted in a new area of process engineering, namely computer-aided molecular design (CAMD) [6]. The original task of the design of molecules is formulated as follows: given a set of desired properties, design a product that satisfies these needs. With the use of CAMD tools, the algorithmic approach to the same problem determines the search place: given a set of available structural groups for the satisfaction of desired properties, formulate a product from these sub-units that satisfies the targets. During the decomposition of a CAMD-based problem, two separate tasks can be derived. As can be seen in *Fig.1*, it can be divided into a forward problem, the prediction of a given property, based on the structural characteristics of a molecule; and a 'reverse' problem, the identification of a molecular structure for the satisfaction of target properties.

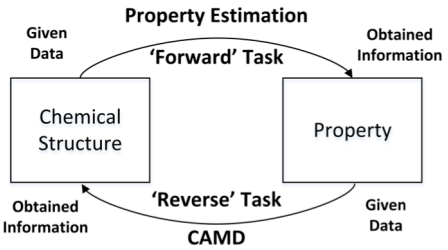


*Figure 1.* Molecular design is a 'reverse' property estimation task.

*Correspondence: janos@abonyilab.com

The forward problem, the estimation of properties, can be carried out by different methods; for example, polymers [1], solvents [7], surfactant solutions [8], refrigerants [9] and ionic liquids [10]. The limitations of any computer-aided molecular design techniques are closely related to the limitations of the property model being used [11]. The prediction of properties can be carried out with numerous types of methods, including group contribution methods (GC), quantitative structure-activity/property relationship methods (QSAR and QSPR), molecular modeling, empirical modeling and correlations, black box models like neural networks (NN) and the combination of these tools [12]. A novel method for property estimation is the COSMO-RS theory published by Klamt *et al*. that combines quantum chemistry and thermodynamics [13].

The 'reverse' problem, the design of candidate molecules with a given set of properties from a set of molecular sub-units is hardly diversified; the existing techniques were developed for specific molecules and applications. The known methods can be divided into two major groups [14].

The huge chemical search space is further complicated by the often competing target properties of the design process. A genetic algorithm is a promising method for the generation of new candidate molecules. Multi-objective optimization algorithms generate a set of optimal solutions. The Pareto fronts of these solutions simultaneously consider several design aspects. Since when solving the problem multiple target properties must be taken into consideration at the same time; the problem has been implemented in a well-established genetic algorithm-based multi-objective optimization environment, the Non-dominated Sorting Genetic Algorithm-II (NSGA-II). The search space is conveniently described by the occurrence of each fragment from a given set of available types of groups, and the "distance" of the properties from the target values is estimated by the Joback method. The feasibility of the molecule is tested by feasibility constraints for branching and the octet rule. These constraints are also used to introduce a special genetic operator that improves the individuals of the populations to ensure the estimation of the properties is based on only reliable structures. Thus as the result, an evolutionary approach for solving molecular design problems with descriptors of varying dimensionality has been developed, that moves effectively towards the Pareto optimal front.

In the present work the definition of the design problem is followed by a theoretical overview of the used property estimation method and of the nature of genetic algorithms paying special attention to NSGA-II. After the description of the different algorithms, proposed for the solution of the design task, the efficiencies of these approaches are examined through several benchmark problems, and the results are discussed extensively to determine improvements in the applicability of these algorithms.

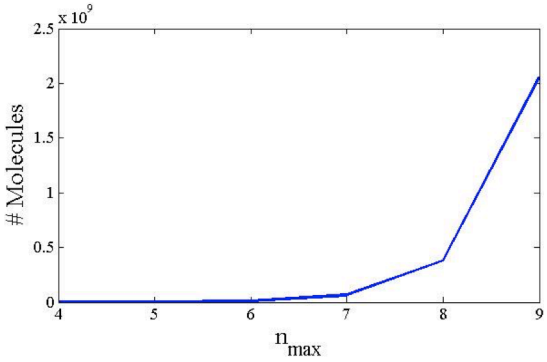


*Figure 2*. The combinatorics of group selection.

## 2. Methodology

### 2.1. Problem Formulation

In the first class, numerous candidate molecules are created randomly from a given set of groups. The number of candidate molecules that can be generated by selecting $N$ groups from a set of $K$ groups, allowing repetition and ignoring the order of selection can be determined by *Eq.(1)*.

$$C^{R}(K,N) = \frac{(K+N-1)!}{N!(K-1)} \quad (1)$$

The total number of candidate molecules that can be selected from a set of $K$ groups is the sum of the results of *Eq.(1)* from $N_1$ until $N_{max}$ as is given by *Eq.(2)* ($N_1$ is practically equal to at least 2, as no molecules consist of only one fragment). The sum of candidates is equal to

$$\sum_{N_1}^{N_{max}} C^{R}(K,N) = \sum_{N_1}^{N_{max}} \frac{(K+N-1)!}{N!(K-1)} \quad (2)$$

According to *Eq.(2)* the total number of candidate molecules can undergo a combinatorial explosion as can be seen in *Fig.2*.

In the second class for the solution of the 'reverse' problem, this increased number of candidates must be eliminated by some objective function that expresses the 'distance' from the target. An example of this approach can be a tree structure to mimic the chain of a molecule and reach feasible structures as was carried out in Refs. [12] and [14].

As in the literature, the proposed methods for the design of molecules are highly diversified. Lin *et al*. studied the design of metal catalysts [15], numerous articles can be found for the design of drugs [5, 16-18], Perdomo *et al*. designed and improved biodiesel fuel blends [3] and Kasat *et al*. summarized the applications of genetic algorithms in polymer science including polymer design [2].

To solve the design problem, several computational strategies have been used. Genetic algorithms (GA) are used in several publications [15, 19, 20], a combination of neural networks and genetic algorithms is used in [21] and linear programming is used in [12, 14].

## 2.2. Theoretical Methodologies

The definition of a chemical product design problem is based on the description of design constraints. A set of properties is specified as constraints with specified values with lower and/or upper boundaries. These properties are the *explicit property constraints* as their values can be determined directly by the application of some model calculation or experimentally. In the case of CAMD problems, explicit constraints are evaluated through property estimation methods, these can be, for example, critical properties, solubility indexes, normal boiling points, etc. However, property estimation methods have been significantly improved, there are products, for example, food, fragrances, health and safety products, and aesthetics that cannot be calculated with the use of these models, as these properties are based on subjective opinions or existing knowledge. In the case of these *implicit property constraints* (e.g. taste, aroma, color and health effects of products) the use of databases or the opinion of the designer can be implemented during the evaluation stage. During the basic CAMD process explicit constraints are taken into consideration, these relate mainly to physical properties, and implicit considerations are taken into account during the selection of available molecule fragments or compounds (e.g. no aromatic compounds are taken into consideration, or no halogens or cyanides are available) [6].

To understand the formulation of the solution to this explicit property constraint-based problem, the following information must be taken into consideration as the input information of the molecule design task (according to [12, 14]):

1. Set $G$ of $N_{max}$ groups of which the designed molecule can be composed
2. The boundaries for the specified properties to be satisfied: $P_{lb}^{j}$'s for the lower boundaries and $P_{ub}^{j}$'s for the upper boundaries, where $j = 1, 2, \ldots, m$, the specified properties
3. The lower ($l_{ll}^{i}$) and upper limits ($l_{ul}^{i}$), for the number of appearances of group $i$ in the designed molecule ($i = 1, 2, \ldots, n$)
4. The property $k$ can be estimated via a property estimation method as function $f^{k}$ ($f = 1, 2, \ldots, m$), in the case of group contribution methods $f^{k}$ can be written as $f^{k}(x_1, x_2, \ldots, x_n)$ (where $x_1, x_2, \ldots, x_n$ are the numbers of group types #1, #2, …, #$n$ respectively).

The problem using the expressions above can be formulated as follows: $i$ groups can be chosen from a given set of molecular subunits ($G$) considering the limits of $l_{ll}^{i}$ and $l_{ul}^{i}$, to find all the possible molecular structures, while the property constraints given in *Eq.(3)* are satisfied (where $j = 1, 2, \ldots, m$).

$$P_{lb}^{j} \leq f^{j}(x_1, x_2, \ldots, x_n) \leq P_{ub}^{j} \qquad (3)$$

During the solution of the above-defined CAMD task, the generation and test method can seem to be inefficient as an enormous number of candidate molecules are created which finally turn out to be infeasible molecular structures.

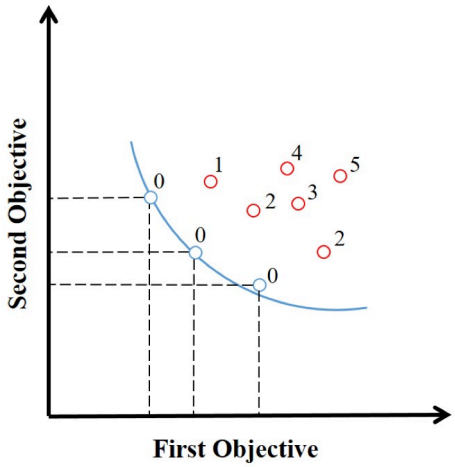


*Figure 3*. A Pareto front (The label of each solution refers to the number of other solutions that dominate them, non-dominating solutions are labeled with zero) [22].

As is familiar among financial and industrial problems, some properties need to be minimized and others maximized between the constraint values, while others need to be close to a specified value. This results in a multiple-objective optimization task with concurring targets. Our purpose is not to find a single solution, but a set of candidate molecules from the Pareto front. Pareto optimal solutions are those for which improvement in terms of one objective can only take place with the worsening of at least one other objective function. Pareto-ranking is the process of determining the rank of each solution through identifying the number of other solutions that dominate it (the number of solutions that are better than it in terms of every objective) [22]. A Pareto front can be seen in *Fig.3*.

In the present work feasibility constraints are implemented in the algorithm to filterout the resultant molecular structures in terms of feasibility. As in this approach the candidates are still filtered out after the property evaluation, the efficiency of the search can still seem to be inefficient. As the solution to this contradiction, a special genetic operator has been introduced that improves the individuals of the populations to ensure the estimation of the properties is based on only reliable structures, and the property evaluation is carried out on solely feasible molecules.

## 2.3. Property Estimation

As the new molecules created in the 'reverse' problem are evaluated via property estimation methods to verify the satisfaction of properties, the success of CAMD tasks depends on, to a large extent, the reliability of the estimation method being used. From the point of view of precision, the highly improved, detailed models seem to be tempting in terms of the application. However, the computational complexity of these models is increased

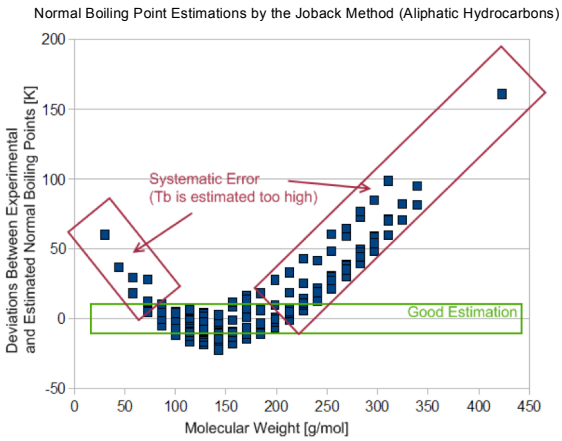


*Figure 4*. Deviations between predicted boiling points and experimental data [24].

as well. In terms of precision and complexity, group contribution methods are promising solutions, as the equation for estimation assumes a linear additivity dependence as presented in *Eq.(4)*, where $\theta_i^j$ is the group contribution value of group $i$ for property $j$, $\theta_0^j$ is the offset value of property $j$ and $P_{est}^j$ is the estimated property value.

$$P_{est}^j(\bar{x}) = \theta_0^j + x_1\theta_1^j + x_2\theta_2^j + \ ... + x_n\theta_n^j \qquad (4)$$

The Joback method, also known as the Joback/Reid method, is proposed to estimate eleven important physical properties of pure materials. During the determination of group contribution values, a common set of structural groups was employed in the regression process. To obtain the minimum values, the minimization of the sum of the absolute errors found from the estimated and the experimental values was carried out. As not the square values, but the absolute values were minimized, the method provides an improved estimation for the majority of the cases, but estimates slightly higher error values for outliers [23]. The systematic deviations of the Joback method in the case of normal boiling points can be seen in *Fig.4*, where experimental data is taken from the Dortmund Data Bank.

The Joback method uses *Eqs.(5-15)* to predict the specific properties as follows:
Normal Boiling Point:

$$T_b[K] = 198 + \sum T_{b,i}x_i \qquad (5)$$

Melting Point:

$$T_m[K] = 122.5 + \sum T_{m,i}x_i \qquad (6)$$

Critical Temperature:

$$T_c[K] = T_b \frac{1}{\left[0.584+0.965\ \sum T_{c,i}x_i - \left(\sum T_{c,i}x_i\right)^2\right]} \qquad (7)$$

Critical Pressure ($N_A$ is the number of atoms in the molecular structure):

$$P_c[\text{bar}] = \left[0.113 + 0.0032 \cdot N_A - \sum P_{c,i}x_i\right]^{-2} \qquad (8)$$

Critical Volume:

$$V_c[\text{cm}^3/\text{mol}] = 17.5 + \sum V_{c,i}x_i \qquad (9)$$

Heat of Formation (ideal gas, 298 K):

$$H^0[\text{kJ/mol}] = 68.29 + \sum H_i^0 x_i \qquad (10)$$

Gibbs Free Energy of Formation (ideal gas, 298 K):

$$G^0[\text{kJ/mol}] = 53.88 + \sum G_i^0 x_i \qquad (11)$$

Heat Capacity (ideal gas, parameters are valid from 273 K to approximately 1000 K):

$$C_P\left[\frac{\text{J}}{\text{molK}}\right] = \sum a_i x_i - 37.93 + [\sum b_i x_i + 0.210] \cdot T + \\ + [\sum c_i x_i - 3.91 \cdot 10^{-4}] \cdot T^2 + \\ + [\sum d_i x_i + 2.06 \cdot 10^{-7}] \cdot T^3 \qquad (12)$$

Heat of vaporization at normal boiling point:

$$\Delta H_{vap}[\text{kJ/mol}] = 15.30 + \sum H_{vap,i}x_i \qquad (13)$$

Heat of Fusion:

$$\Delta H_{fus}[\text{kJ/mol}] = -0.88 + \sum H_{fus,i}x_i \qquad (14)$$

Liquid Dynamic Viscosity ($M_W$ is the molecular weight, the parameters are valid from the melting point up to 0.7 of the critical temperature):

$$\eta_L[\text{Pa} \cdot \text{s}] = M_w \cdot \exp\left(\left[\sum \eta_{a,i}x_i - 597.82\right]/T + \right. \\ \left. + \sum \eta_{b,i}x_i - 11.202\right) \qquad (15)$$

## 2.4. A Promising Approach for the Solution of the Design Task: Genetic Algorithms

Genetic Algorithms (GAs) are stochastic optimization methods that imitate natural selection. GAs provide not a single optimal solution to a problem, but several near-optimal solutions, which is the main advantage of evolutionary algorithms in the field of CAMD, because near-optimal solutions can be further processed later by the designer and the most promising ones can be selected for synthesis.

During the operation of a GA, a population of candidate solutions competes for survival, based on their resemblance to the target values. This resemblance is described by a normalized distance value between 0 and 1 and called the fitness. Candidate molecules are usually described by strings, and the components of these strings represent the 'genes' of the individual. The evaluation of the population is carried out therefore by

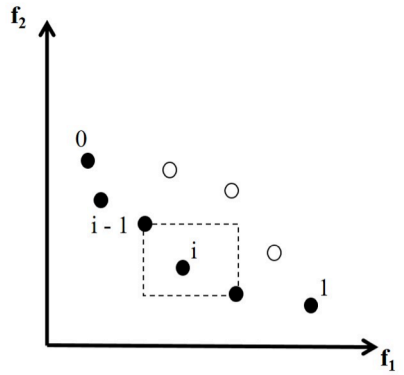


*Figure 5*. The crowding distance of the i[th] solution is the average side length of the cuboid (the front is marked with solid circles) [25].

the calculation of fitness and the surviving members have the chance to reproduce and propagate their genes, thus forming the next generation. This propagation is dependent on the genetic operators applied by the specific algorithm being used, the most common are crossover and mutation. The creation of next generations is continued until convergence is obtained (no considerable improvement is observed), or the maximum number of generations set by the user is reached [21].

### 2.4.1. The Non-dominated Sorting Genetic Algorithm-II (NSGA-II)

For the description of the NSGA-II algorithm, three innovations of the algorithm must be described first: the fast non-dominated sorting procedure, a fast crowded distance estimation method, and the crowded comparison operator [25, 26].

The *fast non-dominated sorting approach* of the NSGA-II is based on the calculation of three entities: the domination count, the number of solution, which dominates the given solution, and the set of solutions that the given solution dominates. In the first non-dominated front the domination count of all solutions is zero. After the determination of the first non-dominated front, each of the solutions dominated by its members is visited and his or her domination count is reduced by one. If the domination count of a solution becomes 0, then it becomes a member of the second non-dominated front. This algorithm is repeated until all fronts are determined.

Along with convergence to the Pareto optimal set of solutions, the maintenance of a healthy spread of solutions is required to avoid the problem of getting stuck in the area of a local Pareto optimum. To prevent this issue, the parameter of *crowding distance* is introduced. During the calculation of this parameter, the average distance between two points on either side of a particular solution along each objective is calculated. The overall crowding distance value is the sum of the individual crowding distance values along each objective. With the use of this parameter, the "density" of solutions in the search place can be calculated. A solution with a higher crowding distance value is less crowded by other solutions, in other words, the outlier solutions can be identified. Thus, the parameter is applicable for the maintenance of diversity. The crowding distance computation for two objectives is illustrated in *Fig.5*.

The goal of the genetic algorithm, to obtain a uniformly spread Pareto-optimal front, is reached with the help of the *crowded-comparison operator*. The operator guides the selection between two possible solutions as follows:

1. If the non-domination ranks of two solutions differ, the solution which dominates the other is preferred, in other words, whose domination index is less.
2. If the non-domination ranks of two solutions are equal (the two solutions are from the same front), then the solution of the less crowded region is preferred.

The main loop of the NSGA-II can be explained by understanding the operators described above. The randomly created initial parent population (with $N$ members) is sorted based on the non-domination rank. The crossover *Eqs.(16-17)* and mutation *Eqs.(18-19)* operators are applied to create the next generation (with $N$ members) [27]. The algorithm applies to the intermediate crossover, which creates two children from two parents: *parent1* and *parent2* (*child* and *parent* are vectors, $\bar{x}$, containing the results of the specific problems),

$$child1 = parent1 + rand \cdot ratio \cdot (parent2 - -parent1) \quad (16)$$

$$child2 = parent2 - rand \cdot ratio \cdot (parent2 - -parent1) \quad (17)$$

where *ratio* is a scalar between 0 and 1, and *rand* stands for a random number,

The applied Gaussian mutation adds a normally distributed random number to each variable,

$$child = parent + S \cdot rand \cdot (ub - lb) \quad (18)$$

$$S = scale \cdot \left(1 - shrink \cdot \frac{currGen}{maxGen}\right) \quad (19)$$

where *scale* is a scalar, that determines the standard deviation of the random number generated and *shrink* is a scalar between 0 and 1. As the optimization progresses, this shrink parameter decreases the mutation range. *currGen* and *maxGen* are the numbers of the current and maximal generations, respectively.

Since elitism is introduced, the creation of the first population differs from the creation of a subsequent one. The algorithm is described for the $t^{th}$ generation.

A combined population ($R_t$) (with $2N$ members) is created by the summation of the parent population ($P_t$) and the population obtained by the use of crossover and mutation operators ($Q_t$). The population $R_t$ is sorted according to non-domination and as all previous and current population members are included, elitism is ensured. Now solutions belonging to the first non-

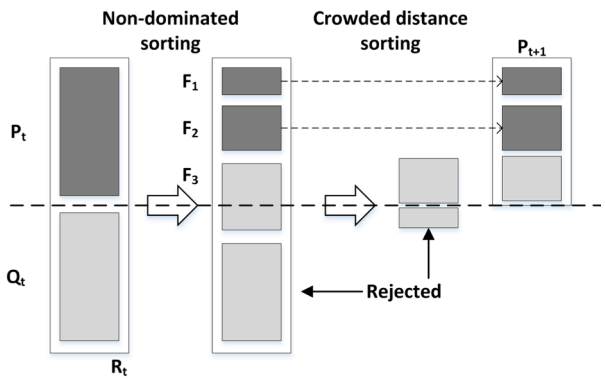


*Figure 6*. Graphical illustration of the NSGA-II procedure [25].

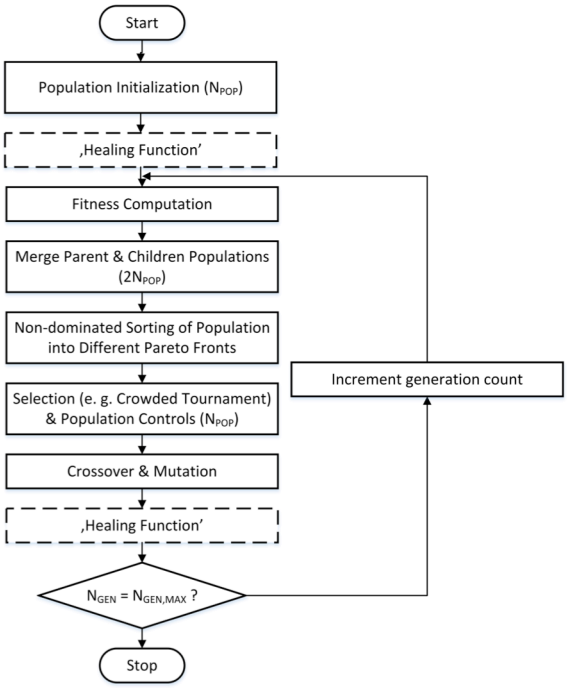


*Figure 8*. The algorithm of the genetic algorithm (the steps in dashed box are only implemented in Algorithm 5).

dominated front ($F_1$) are chosen for the next generation ($P_{t+1}$) (if the size of $F_1$ is smaller than $N$). This selection for the next generation is continued until the number of members from $F_1$ to $F_i$ is larger than $N$. In these cases $F_i$ is sorted based on the crowded-comparison operator and the best solutions are chosen to fill the empty slots of the new population. The NSGA-II procedure is illustrated in *Fig.6*.

## *2.5. Description of the Proposed Algorithm*

The developed algorithm therefore needs to solve effectively the CAMD tasks based on the needs of the industrial and research work. By taking into consideration these needs, the optimal properties of the desired molecules can be defined, and this chemical information is essential for the defining of input parameters for the genetic algorithm. This schematic algorithm of the design process can be seen in *Fig.7*.

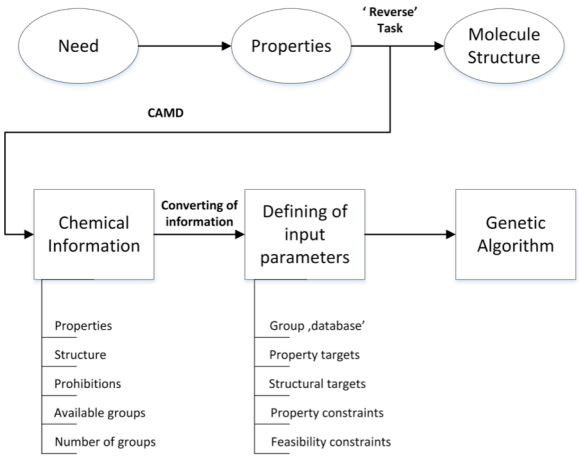


*Figure 7*. The algorithm of the design process.

The standard genetic algorithm was modified according to the task of the design of molecules. The chemical information is converted into input parameters. Thus, the type, minimum and a maximum number of available groups, the property and structural targets (e.g. acyclic, monocyclic, bicyclic structures), and the property and feasibility constraints (octet rule and a rule for branching) are defined. The algorithm of NSGA-II can be seen in *Fig.8*.

The efficiency of five different types of algorithm was tested in the present work. All the algorithms were implemented in MATLAB.

### 2.5.1. The 'base case' Algorithm 1

Only the type, the minimum and a maximum number of available groups, the target properties and the property constraints are defined. No structural targets or feasibility evaluations are implemented.

### 2.5.2. The 'octet rule' Algorithm 2

Besides the objectives and constraints of Algorithm 1, the octet rule is defined as target parameter. The octet rule is described in *Eq.(20)*,

$$\sum_i (2 - v_i) x_i = 2m \qquad (20)$$

where $x_i$, $v_i$ are the number and valency, respectively, of groups of type $i$ and $m$ = 1, 0 or -1 for acyclic, monocyclic and bicyclic groups, respectively [28]. The valency parameter in this context means the number of available bonds on a group (thus the valency of double bonds counts as a single valency in this context).

### 2.5.3. The 'octet rule as a soft constraint' Algorithm 3

Besides the structure of Algorithm 2, the algorithm contains the octet rule (*Eq.(20)*) as a soft constraint to aspire the program to reach a feasible structure according to the octet rule. The soft constraint was defined as a curve similar to a reverse Gaussian distribution according to *Eq.(21)*, where $Res.Oct.$ is the result of the octet rule reordered to give 0 when the constraint is satisfied.

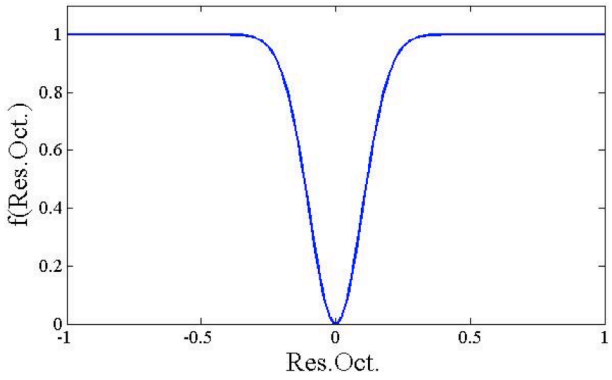


*Figure 9*. The curve of the used soft constraint.

$$f(Res.Oct.) = w \cdot (1 - e^{-\frac{(Res.Oct.)^2}{2\sigma^2}}) \qquad (21)$$

As can be seen in *Fig.9*, the curve of this function has a descending value near the $Res.Oct. = 0$ value, and is equal to 0 at exactly $x = 0$. The parameter $w$ shows the 'weight', the constant value far from $Res.Oct. = 0$, and $\sigma$ stands for the 'width' (thus the 'sharpness') of the function. A sharp function type ($w = 1$, $\sigma = 0.1$) was chosen as structural feasibility is not satisfied with mild ones.

#### 2.5.4. The 'octet and branching rule' Algorithm 4

Only the octet rule cannot describe the structural feasibility. Two adjacent groups cannot be linked by more than one bond. The valency parameter in this context still means the number of available bonds on a group, as defined in the description of Algorithm 2. Given $x_j$ groups of type $j$ with valency $v_j$, a total of $x_j(v_j - 2) + 2$ attachments are available for bonding [28].

$$\sum_{i \neq j} x_i \geq x_j(v_j - 2) + 2 \qquad (22)$$

$$\sum_{i} x_i \geq x_j(v_j - 1) + 2 \qquad (23)$$

#### 2.5.5. The 'healing function' Algorithm 5

The genetic algorithm is further improved with the introduction of a novel genetic operator called the healing function. The function is implemented in the algorithm at two points, after the initialization of the first population and after the selection, mutation and crossover steps are conducted in each generation. This function serves the population entities to be feasible with the help of the branching rule described in *Eq.(23)*. If the equation is not fulfilled then the difference from the optimal value shows the bonds needed in the molecule to reach the feasible structure. Using this value, the algorithm chooses as many available groups with one bond as needed for feasibility and completes the molecule with these groups. During the definition of the input parameters the number of available groups with one bond in the molecule is increased to ensure the availability of these groups. The number of branches and the groups with 3 or more bonds, must be significantly less than the ones with one bond available, e.g. 4 and 20 respectively, thus the termination of every chain in the molecule is ensured.

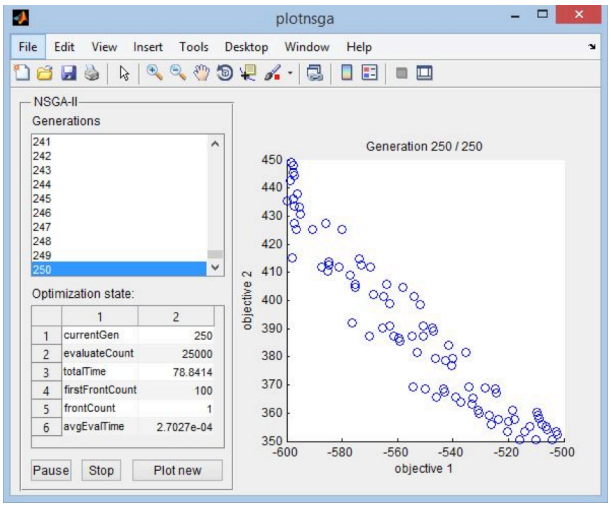


*Figure 10*. The plot window of NSGA-II.

### 2.6. The Evaluation of the Results

As the purpose of the current work is the development of an effective algorithm for the design of molecules obtaining target properties, the comparison of the results is an essential task to check the improvement. This efficiency inspection is carried out over two steps.

First a visual evaluation was carried out as the MATLAB implementation of the NSGA-II algorithm plots every generation as the script runs. The plot window of NSGA-II can be seen in *Fig.10*. Then the number of solutions is counted and, as the feasible structure is not ensured in every algorithm, the feasible ones based on the octet and branching rules are counted as well. The last generation in which the population of the Pareto front was changed was also determined.

## 3. Results and Discussion

The effect of the various user determined input parameters was examined and presented in several benchmark problems for the identification of different chemicals having the desired physical and chemical properties, as estimated by the multi-dimensional property model. The genetic algorithm worked with 250 generations containing 100 population members each. The effectivity of the different algorithms is compared through these design tasks.

### 3.1. Checking the Effectivity through the Search for Predefined Molecules

First the effectivity and applicability of the different algorithms were tested in terms of the search for different, predefined molecules. The properties of simple molecules were calculated *via* the Joback method and these values were set as targets to avoid the inaccuracy of the estimation method. The property constraints were set around these target values as given

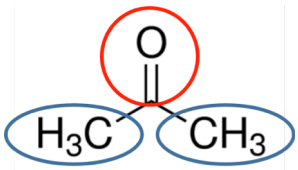


*Figure 11*. The structure of acetone fragmented according to the groups of the Joback method.

*Table 1*. The boiling and melting points of acetone.

| | Experimental values | Estimated values |
|---|---|---|
| $T_m$ [K] | 178.25 | 173.50 |
| $T_b$ [K] | 329.45 | 321.91 |

*Table 2*. The results of the search for acetone.

| Algorithm | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Solutions | 2 | 2 | 1 | 1 | 2 |
| (feasible) | (1) | (1) | (0) | (1) | (2) |
| $N_{gen,change}$ | 95 | 10 | 8 | 9 | 127 |
| acetone | yes | yes | - | yes | - |

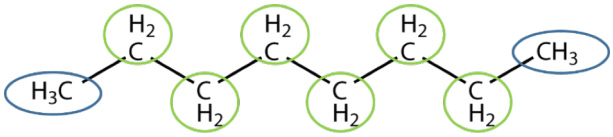


Figure 12. The structure of *n*-octane fragmented according to the groups of the Joback method.

Table 3. The boiling and melting points of octane.

| | Experimental values | Estimated values |
|---|---|---|
| $T_m$ [K] | 216 | 179.92 |
| $T_b$ [K] | 398 | 382.44 |

*Table 4*. The results of the search for *n*-octane.

| Algorithm | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Solutions | 10 | 5 | 1 | 1 | 5 |
| (feasible) | (2) | (1) | (0) | (1) | (5) |
| $N_{gen,change}$ | 191 | 58 | 17 | 207 | 47 |
| n-octane | - | yes | - | yes | - |

below. The search for two simple molecules, acetone, as its structure is quite simple, and *n*-octane were carried out to validate the algorithms.

### 3.1.1. Acetone

The structure of acetone can be represented by two methyl (blue circles) and one ketone (red circle) groups using the groups of the Joback method as can be seen in *Fig. 11*. The experimental and estimated boiling and melting points are presented in *Table 1*.

The input parameters for the design task were as follows:

*Available groups*:
$-CH_3$, $-CH_2-$, >CH-, >C<, -F and >C=O

*Number of available groups*:
0–4 for all the available groups
(in the case of Algorithm 5, the $-CH_3$, and -F numbers were set to 20 to allow healing)

*Target properties*:
$T_m$ = 173.50 K, $T_b$ = 321.91 K (estimated values)

*Property constraints*:
150 K < $T_m$ < 200 K, 300 K < $T_b$ < 350 K

*Target molecule structure*:
acyclic.

The results of different algorithms are presented in *Table 2*. Next to the number of solutions, the number of feasible solutions is presented in brackets and the value $N_{gen,change}$ stands for the last generation that changed the population of the Pareto front. The last row of the table shows if acetone is among the resultant structures.

Algorithm 5 seemed to be less effective in the light of the unsuccessful search for the structure of acetone, but if we consider that this algorithm could use up to 20 pieces of $-CH_3$ and -F groups, we can understand that the search place is significantly bigger than in the case of Algorithms 1-4.

### 3.1.2. *n*-Octane

The structure of *n*-octane can be represented by 2 methyl (blue circles) and 6 methylene (green circles) groups using the groups of the Joback method as can be seen in *Fig.12*. The experimental and estimated boiling and melting points are presented in *Table 3*.

The input parameters for the design task were as follows:

*Available groups*:
$-CH_3$, $-CH_2-$, >CH-, >C<, -F and >C=O

*Number of available groups*:
0–6 for all the available groups
(in the case of Algorithm 5, the $-CH_3$, and -F numbers were set to 20 to allow healing)

*Target properties*:
$T_m$ = 179.92 K, $T_b$ = 382.44 K (estimated values)

*Property constraints*:
150 K < $T_m$ < 250 K, 350 K < $T_b$ < 450 K

*Target molecule structure:*
acyclic.

The results of different algorithms are presented in *Table 4*, where the last row shows if *n*-octane is among the resultant structures. As in the case of the search for acetone, Algorithm 5 had a significantly bigger search place than Algorithms 1-4; thus, it could not find the structure of *n*-octane.

*Table 5*. The results of Case Study 1.

| Algorithm | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Solutions (feasible) | 89 (14) | 88 (14) | 1 (0) | 2 (2) | 42 (42) |
| $N_{gen,change}$ | 250 | 250 | 219 | 186 | 250 |

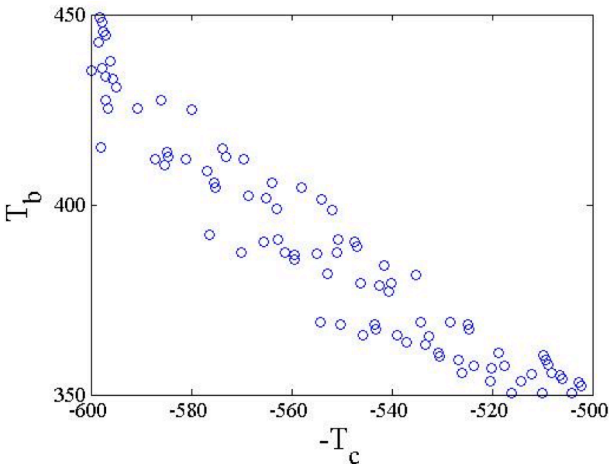


*Figure 13*. The Pareto Front of Algorithm 1 in Case Study 1.

## 3.2. Examples for Testing the Algorithms

In the following examples the effectivity of the proposed algorithms is tested via two benchmark problems. As concurring minimum and maximum search objectives are set as target variables, a Pareto front can be obtained, which represents the applicability of the genetic algorithms to solve CAMD tasks.

### 3.2.1. Case Study 1.

The input parameters for the design task were as follows:

*Available groups*:
$-CH_3$, $-CH_2-$, >CH-, >C<, -F and >C=O

*Number of available groups*:
0–4 for all the available groups
(in the case of Algorithm 5, the $-CH_3$ and -F numbers were set to 20 to allow healing)

*Target properties*:
$\max(T_c)$, $\min(T_b, T_m)$

*Property constraints*:
$500\ K < T_c < 600\ K$
$350\ K < T_b < 450\ K$
$100\ K < T_m < 200\ K$

*Target molecule structure*:
acyclic.

The results of Case Study 1 can be seen in *Table 5*. As the number of the last generation which changed the composition of the population is relatively high (250 is the maximum number of generations), the algorithms seem to evolve effectively towards the Pareto front. In the case of Algorithms 1 and 2, many of the found solutions proved to be infeasible, as expected since no feasibility constraints were involved in Algorithm 1 and no constraint for branching was available in Algorithm 2. The soft constraint of Algorithm 3 seemed to be ineffective. Algorithm 4 found only 2 feasible solutions, but these solutions were all feasible ones and they were close to the target properties. Algorithm 5 was very effective, although the search place was significantly bigger than in the case of Algorithms 1-4. The Pareto front of Algorithm 1 can be seen in *Figure 13*.

*Table 6*. The results of Case Study 2.

| Algorithm | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Solutions (feasible) | 75 (0) | 95 (5) | 2 (0) | 2 (2) | 74 (74) |
| $N_{gen,change}$ | 250 | 250 | 223 | 197 | 250 |

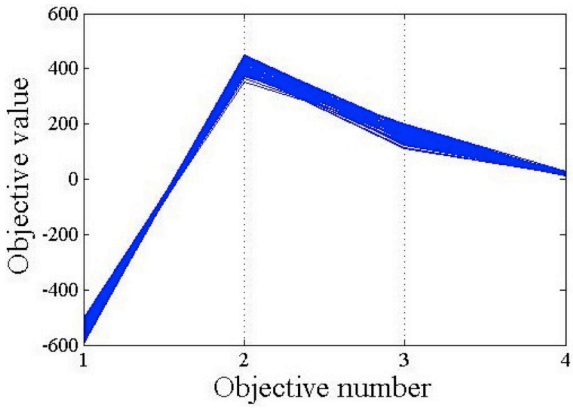


*Figure 14*. The results of Algorithm 5 in Case Study 2. Candidate molecules are represented by blue lines.

### 3.2.2. Case Study 2.

The input parameters for the design task were as follows:

*Available groups*:
$-CH_3$, $-CH_2-$, >CH-, >C<, -F, -Cl, -Br, -I, -OH, >C=O

*Number of available groups*:
0–6 for all the available groups
(in the case of Algorithm 5, the $-CH_3$, -F, -Cl, -Br, -I, and -OH numbers were set to 20 to allow healing)

*Target properties*:
$\max(T_c)$, $\min(T_b, T_m, P_c)$

*Property constraints*:
$500\ K < T_c < 600\ K$
$350\ K < T_b < 450\ K$
$100\ K < T_m < 200\ K$
$10\ bar < P_c < 30\ bar$

*Target molecule structure*:
acyclic.

In the case of Case Study 2 (*Table 6*), a wider searching range was available for the design task: 10 types of available groups with a maximum of six pieces of each (except for Algorithm 5, see the targets and constraints) and another target property was set, the minimization of critical pressure between the property constraints. The results were similar to Case Study 1, Algorithms 1 and 2 fundnd several infeasible results with the appropriate properties; Algorithm 3 seemed to be ineffective, Algorithms 4 and 5 provided reliable results, although Algorithm 5 still found more results according to the more pieces of available chain-terminating groups (groups with 1 valency). As can be seen from the results the improvement of the algorithms increases the number of feasible solutions significantly. The results of Algorithm 5 in Case Study 2 can be seen in *Figure 14*.

## 4. Conclusion

The design of molecules with specified properties has an increasing importance in the modern chemical industry. We proposed a multi-objective evolutionary optimization-based approach to take into account several objectives and constraints (e.g. financial aspects, toxicity). The algorithms generate a set of molecules arranged in a Pareto front related to the conflicting design targets calculated by the Joback method. To get reliable molecular structures we defined soft constraints based on the octet rule. The branching of the molecules was also tested and a healing function was designed to provide reliable results. The application of the proposed algorithms can be useful in the industry.

In the future, we are going to improve the algorithms with problem-specific genetic operators. These modifications seem to be promising for the significant increase in the search efficiency.

## SYMBOLS

| | |
|---|---|
| $C^{\mathrm{R}}(K,N)$ | selection of $N$ groups from a set of $K$ groups |
| $G$ | the set of $n$ groups from which the designed molecule can be composed |
| $n_{\max}$ | the available groups of the specified design task |
| $x$ | the number of appearances of the specified group |
| $P$ | a specified property value |
| $P_{\mathrm{est}}$ | estimated property value |
| $P_{\mathrm{exp}}$ | experimental property value |
| $P_{\mathrm{lb}}$ | the lower boundary of the specified property value |
| $P_{\mathrm{ub}}$ | the upper boundary of the specified property value |
| $l_{\mathrm{ll}}$ | the lower limit for the number of appearances of the specified group |
| $l_{\mathrm{ul}}$ | the upper limit for the number of appearances of the specified group |
| $x_1, \ldots, x_n$ | the number of group type #1, …, #*n*, resp. |
| $a_i^j$ | is the group contribution value of group *i* for property *j* |
| $T_{\mathrm{b}}$ | normal boiling point |
| $T_{\mathrm{m}}$ | normal melting point |
| $T_{\mathrm{c}}$ | critical temperature |
| $P_{\mathrm{c}}$ | critical pressure |
| $V_{\mathrm{c}}$ | critical volume |
| $H^0$ | heat of formation |
| $G^0$ | Gibbs free energy of formation |
| $C_{\mathrm{P}}$ | heat capacity |
| $\Delta H_{\mathrm{vap}}$ | heat of vaporization |
| $\Delta H_{fus}$ | heat of fusion |
| $\eta_{\mathrm{L}}$ | liquid dynamic viscosity |
| *ratio* | a scalar between 0 and 1 |
| *rand* | a generated random number |
| *currGen* | the number of the current generation |
| *maxGen* | the number of the maximal generation |
| *scale* | a scalar, that determines the standard deviation of the random number generated |
| *shrink* | scalar between 0 and 1. As the optimization progresses this parameter decreases the mutation range |
| $R_t$ | A combined population (with 2*N* members) |
| $P_t$ | parent population |
| $Q_t$ | the population obtained by the use of crossover and mutation operators |
| $F$ | the non-dominated front |
| $v$ | the valency of the specified group. The valency parameter in this context means the number of available bonds on a group (thus the valency of double bonds counts as a single valency in this context). |
| $m$ | the type of the designed molecule ($m = 1, 0$ or -1 for acyclic, monocyclic and bicyclic groups, respectively) |
| $w$ | the 'weight' of the soft constraint curve (the constant value far from $x = 0$) |
| $\sigma$ | the 'width' (the 'sharpness') of the soft constraint function curve |

## Acknowledgement

The research of J.A. has been supported by the National Research, Development and Innovation Office (NKFIH), project # OTKA 116674 entitled "Process mining and deep learning in the natural sciences and process development". G.D. was supported by the "A Pannon Egyetem tudományos műhelyeinek támogatása" TÁMOP-4.2.2.B-15/1/KONV-2015-0004 project.

## REFERENCES

[1] Camarda, K.V.; Maranas, C.D.: Optimization in polymer design using connectivity indices, *Ind. Engng. Chem. Res.*, 1999 **38**(5), 1884–1892 DOI: 10.1021/ie980682n

[2] Kasat, R.B.; Ray, A.K.; Gupta, S.K.: Applications of genetic algorithms in polymer science and engineering, *Mat. Manufact. Proc.*, 2003 **18**(3), 523–532 DOI: 10.1081/AMP-120022026

[3] Perdomo, F.A.; Perdomo, L.; Millán, B.M.; Aragón, J.L.: Design and improvement of biodiesel fuel blends by optimization of their molecular structures and compositions, *Chem. Engng. Res. Design*, 2014 **92**(8), 1482–1494 DOI: 10.1016/j.cherd.2014.02.011

[4] Joback, K.G.: Computer aided molecular design (CAMD): Designing better chemical products. (Molecular Knowledge Systems Inc., Bedford, NH U.S.A.) 1998-2016 www.molecularknowledge.com

[5] Schneider, G.; Hartenfeller, M.; Reutlinger, M.; Tanrikulu, Y.; Proschak, E.; Schneider, P.: Voyages to the (un)known: Adaptive design of bioactive compounds, *Trends Biotechn.*, 2009 **27**(1), 18–26 DOI: 10.1016/j.tibtech.2008.09.005

[6] Gani, R.; Achenie, L.E.K.; Venkatasubramanian, V.: Introduction to CAMD in computer aided chemical engineering (Eds.: Luke, R.G.; Achenie, L.E.K.; Venkat, V.: Elsevier, Amsterdam, The Netherlands), 2002 Chapter 1, pp. 3–21 DOI: 10.1016/S1570-7946(03)80003-2

[7] Gani, R.; Jiménez-González, C.; Constable, D.J.C.: Method for selection of solvents for promotion of organic reactions, *Comp. Chem. Engng.*, 2005 **29**(7), 1661–1676 DOI: 10.1016/j.compchemeng.2005.02.021

[8] Camarda, K.V.; Bonnell, B. W., Maranas, C. D., Nagarajan, R.: Design of surfactant solutions with optimal macroscopic properties, *Comp. Chem. Engng.*, 1999 **23**(Supplement), S467–S470 DOI: 10.1016/S0098-1354(99)80115-X

[9] Sahinidis, N.V.; Tawarmalani, M.; Yu, M.: Design of alternative refrigerants *via* global optimization, *AIChE J.*, 2003 **49**(7), 1761–1775 DOI: 10.1002/aic.690490714

[10] McLeese, S.E.; Eslick, J.C.; Hoffmann, N.J.; Scurto, A.M.; Camarda, K.V.: Design of ionic liquids *via* computational molecular design, *Comp. Chem. Engng.*, 2010 **34**(9), 1476–1480 DOI: 10.1016/j.compchemeng.2010.02.017

[11] Gani, R.: Computer-aided methods and tools for chemical product design, *Chem. Engng. Res. Design*, 2004 **82**(11), 1494–1504 DOI: 10.1205/cerd.82.11.1494.52032

[12] Holenda, B.; Holenda, B.; Dallos, A.; Nagy, Á.; Friedler, F.; Fan, L.-T.: A combinatorial approach for generating environmentally benign solvents and separation agents, *Chem. Eng. Trans., Ser.*, 2003 **3**, 871–875

[13] Klamt, A.: Conductor-like screening model for real solvents: A new approach to the quantitative calculation of solvation phenomena, *J. Phys. Chem.*, 1995 **99**(7), 2224–2235 DOI: 10.1021/j100007a062

[14] Friedler, F.; Fan, L.T.; Kalotai, L.; Dallos, A.: A combinatorial approach for generating candidate molecules with desired properties based on group contribution, *Comp. Chem. Engng.*, 1998 **22**(6), 809–817 DOI:10.1016/S0098-1354(97)00253-6

[15] Lin, B.; Chavali, S.; Camarda, K.; Miller, D.C.: Computer-aided molecular design using Tabu search, *Comp. Chem. Engng.*, 2005 **29**(2), 337–347 DOI: 10.1016/j.compchemeng.2004.10.008

[16] Soto, A.J.; Cecchini, R.L.; Vazquez, G.E.; Ponzoni, I.: Multi-objective feature selection in QSAR using a machine learning approach, *QSAR & Comb. Sci.*, 2009 **28**(11–12), 1509–1523 DOI: 10.1002/qsar.200960053

[17] Hii, C.E.A.: Evolving toxicity models using multigene symbolic regression and multiple objectives, *Int. J. Mach. Learn. Comp.*, 2011 **1**, 30–35 DOI: 10.7763/IJMLC.2011.V1.5

[18] Manoharan, P.E.A.: Rationalizing fragment-based drug discovery for BACE1: Insights from FB-QSAR, FB-QSSR, multi-objective-QSPR, and MIF studies, *J. Comput. Aided Mol. Des.,* 2010 **24**, 843–864 DOI: 10.1007/s10822-010-9378-9

[19] Herring III, R.H.; Eden, M.R.: Evolutionary algorithm for *de novo* molecular design with multi-dimensional constraints, *Comp. Chem. Engng.*, 2015 **83**, 267–277 DOI: 10.1016/j.compchemeng.2015.06.012

[20] Weber, L.: Evolutionary combinatorial chemistry: application of genetic algorithms, *Drug Discovery Today*, 1998 **3**(8), 379–385 DOI: 10.1016/S1359-6446(98)01219-7

[21] Venkatasubramanian, V.; Sundaram, A.; Chan, K.; Caruthers, J.M.: Computer-aided molecular design using neural networks and genetic algorithms. Genetic algorithms in molecular modeling (Ed.: Devillers, J.: Academic Press, London, UK) 1996 DOI: 10.1016/B978-012213810-2/50012-8

[22] Nicolaou, C.A.; Brown, N.: Multi-objective optimization methods in drug design, *Drug Discovery Today: Technologies*, 2013 **10**(3), e427–e435 DOI: 10.1016/j.ddtec.2013.02.001

[23] Joback, K.G.; Reid, R.C.: Estimation of pure-component properties from group-contributions, *Chem. Engng. Commun.*, 1987 **57**(1–6), 233–243 DOI: 10.1080/00986448708960487

[24] Shin Hyo Bang, S.J.L.; Taeseon Y.: Boiling point estimation program especially for aromatic compounds supplementing the Joback method, *Int. J. Chem. Engng. Appl.*, 2014 **5**(4), 331–334 DOI: 10.7763/IJCEA.2014.V5.404

[25] Deb, K.; Agrawal, S.; Pratap, A.; Meyarivan, T.: A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II, in Parallel problem solving from nature PPSN (Eds.: Schoenauer VI, M.: Deb, K.; Rudolph, G.; Yao, X.; Lutton, E.; Merelo, J. J.; Schwefel, H.-P., Springer, Berlin, Germany) 2000 pp. 849–858 DOI: 10.1007/3-540-45356-3_83

[26] Deb, K.: Multi-objective genetic algorithms: problem difficulties and construction of test problems, *Evolut. Comp.,* 1999 **7**, 205–230 DOI: 10.17877/DE290R-5636

[27] Song, L.: NGPM - A NSGA-II Program in MATLAB, User Manual, 2011 www.mathworks.com/matlabcentral/fileexchange/31166-ngpm-a-nsga-ii-program-in-matlab-v1-4

[28] Odele, O.; Macchietto, S.: Computer aided molecular design: A novel method for optimal solvent selection, *Fluid Phase Equil.*, 1993 **82**, 47–54 DOI: 10.1016/0378-3812(93)87127-M