Systems & Control Transactions

Research Article - Peer Reviewed Conference Proceeding
ESCAPE 35 - European Symposium on Computer Aided Process Engineering
Ghent, Belgium. 6-9 July 2025
Jan F.M. Van Impe, Grégoire Léonard, Satyajeet S. Bhonsale,
Monika E. Polańska, Filip Logist (Eds.)

PSE PRESS

Computer-Aided Molecular Design for Bio-Based Solvent Selection from Citrus and Coffee Wastes for Furfural Extraction

Giovana C. A. Netto[a*], Moisés Teles dos Santos[a], and Vincent Gerbaud[b]

[a] University of São Paulo (USP), Department of Chemical Engineering, São Paulo, São Paulo, Brazil
[b] Université de Toulouse, CNRS, INP, UPS, Laboratoire de Génie Chimique (LGC), Toulouse, France
* Corresponding Author: giovana.netto@usp.br.

ABSTRACT

The global reliance on fossil-based solvents has driven the search for sustainable alternatives. This study employs the IBSS® CAMD tool to evaluate building blocks derived, directly or indirectly, from agricultural residues - specifically orange and coffee wastes-, to replace toluene in furfural extraction. A three-stage methodology was implemented: (1) identification of potential building blocks from residues, (2) multi-objective optimization using genetic algorithms and group contribution models for properties calculation, and (3) analysis of the resulting candidates based on performance indicators. A total of 13 families were evaluated, generating millions of candidates. Target properties included minimization of Hansen Solubility Parameters (HSP) distance, boiling point above 250°C, melting point below 10°C, flash point above 61°C, and octanol-water partition coefficient (log(kow)) below 3. The most promising candidates were derivatives of glycerol (performance: 0.9986), limonene (0.9937), and furfural (0.9931), which exhibited low bioaccumulation potential, non-flammability, and liquid phases at typical usage temperatures. Notably, all the top-performing molecules derived from these building blocks outperformed toluene (performance: 0.6724) in terms of the selected properties. The findings highlight the dual benefits of addressing agricultural waste management challenges while advancing sustainable industrial practices. Further studies are required to confirm their stability and scalability for industrial applications.

Keywords: Biomass, Genetic Algorithm, CAMD, Furfural, Solvent, Agricultural Wastes, Molecular Design

INTRODUCTION

Brazil has an outstanding agricultural production that reaches 94 million ha [1], which represents 11% of the nation's total territory. The country stands as the world's leading producer of several key commodities, including coffee and oranges. However, alongside such intensive production comes an increasing concern about the sustainable management of the vast amounts of agricultural waste generated throughout the supply chain [2].

In 2023, Brazil's coffee production reached 3.5 million metric tons [3], primarily through the dry processing method, where coffee cherries are dried immediately after harvest [4]. Notably, 90–99% of the fruit mass is classified as waste [5]. Also, in 2023, Brazil produced 18 million metric tons of orange, of which 72% [6] was destinated for orange juice production. During this process, 50–60% of the fruit mass becomes waste, primarily composed of seeds and peel [7].

Agricultural wastes from coffee and orange production are typically disposed in fields, repurposed as animal feed or incinerated [7]. These practices contribute to environmental impacts and these approaches overlook the significant potential of these residues. Rich in bioactive compounds, such as antioxidants and fibers, coffee and orange byproducts offer untapped opportunities for sustainable valorization [2].

The residues contain building blocks that can be extracted or produced from them and one that stands out is furfural. Furfural is produced via the dehydration of xylose, a sugar present in the hemicellulose fraction of

lignocellulosic biomass. It has a wide range of industrial applications, including the production of plastics, pharmaceuticals, agrochemical products, non-petroleum-derived chemicals and it is a precursor for various furan-based chemicals and feedstocks [8].

Although the synthesis of furfural dates to 1921, the process still faces challenges, including low efficiency (~50%) and high energy demands, especially when distillation is employed as a purification method. Therefore, liquid liquid extraction (LLE) appears as a possible purification method and the main solvent used is toluene, a fossil-based toxic solvent [9].

The procedure for solvent selection plays a crucial role in determining the efficiency and viability of the proposed new solvent. In this study, Computer Aided Molecular Design (CAMD) was chosen as an effective early-stage tool for discovering new molecules, which leverages the simplicity of semi-empirical quantitative structure-property relationships (QSPRs) in conjunction with genetic optimization algorithms [10].

An in-house CAMD tool called IBSS® was employed. This tool was developed in compliance with a model driven software engineering approach, and it is based on different analysis models (UML and BPMN) and object-oriented implementation models dedicated to Microsoft.NET framework (C# and VB.NET) [11]. IBSS® has been successfully applied in various contexts, such as identifying solvents for the perfume industry [12], developing greener solvents for photovoltaic panel production [13], and substituting hexane in metabolite extraction from biomass feedstock [14].

The objective of this work was to evaluate solvents derived from coffee and orange residues, to extract furfural from an aqueous solution with methanol and acetic acid, in substitution to toluene.

METHODOLOGY AND PROBLEM FORMULATION

To accomplish the goal of finding a suitable bio-based solvent to replace toluene for furfural extraction, a three-stage methodology, depicted in Figure 1, was proposed based on the work of Heintz *et al.* [15].

Molecular Fragments Literature Review

The first stage entailed a literature review to identify potential building blocks derived, directly or indirectly, from coffee and orange residues. From an initial pool of 46 building block families, 13 were selected for further evaluation: furoate, geranyl, glucaric acid, glutamic acid, hydroxymethylfurfural, hydroxypropionic acid, levulinic acid, limonene, 5-methylfurfural, oleic acid, succinic acid, glycerol, and furfural itself. Considering the derivatives from the abovementioned building block families, 103 different fixed fragments were analyzed.

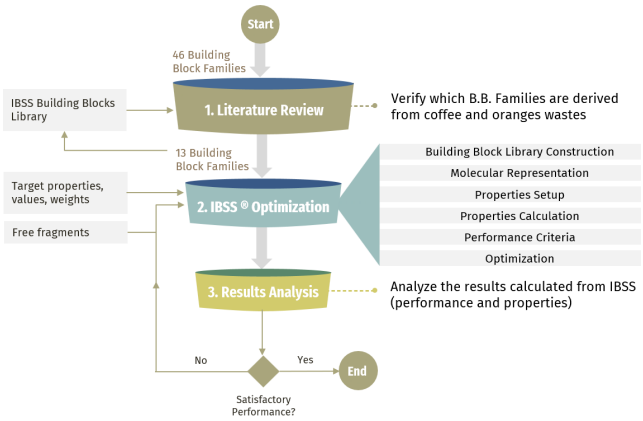


Figure 1: Methodology for solvent selection using IBSS® tool.

IBSS® Optimization

The second stage corresponded to the IBSS® optimization, which involves essentially six key components: (1) building block library construction, (2) molecular representation, (3) properties setup, (4) properties calculation, (5) performance criteria for molecule evaluation, and (6) optimization [15]. The output consisted of a .txt file generated for the 200th generation of each fixed fragment studied, containing the calculated properties, performance metrics, and molecular graph structures of the individuals, ranked based on their performance.

Building block library construction

After the literature review, a preliminary verification was conducted to determine whether the molecules selected were already present in the IBSS® library. For those not included, new entries were created by assigning an ID, constructing the molecular graph, and defining functionalization points.

The newly fixed building blocks were drawn, and their SMILES representation were determined, which was entered in “evaluation mode” in the tool, to determine the molecular graph.

This process resulted in the creation of 7 new families of building blocks, comprising a total of 39 new fixed building blocks molecules.

Molecular representation

The selected molecular representation model is the molecular graph. In this representation, the diagonal elements represent non-hydrogen atoms, and the off-diagonal elements represent chemical connections. One of the implemented molecules is represented in Figure 2 as an example.

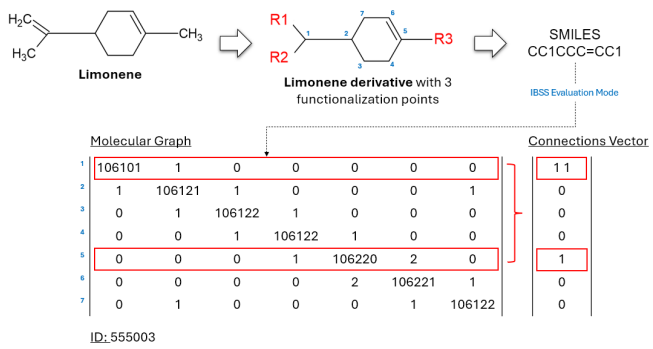


Figure 2: Molecular representation of Limonene.

Properties set up

IBSS® tool allows users to specify the properties desired to calculate the performance, assigning their weights, target values, tolerances, and value of performance in the tolerance value.

The target properties chosen, detailed in Table 1, were defined based on the recommendations of Heintz *et al.* [15]. And the tolerance value reflects the confidence in the model prediction value, which can be assessed with help of the model accuracy and/or the property measurement standard deviation [16]. There were also included the standard deviation for the predictions.

The most critical property is the Hansen Solubility Parameter (HSP) distance, which accounts for three key interaction forces: D (London dispersion forces), P (dipole-dipole polarity interactions), and H (hydrogen bonding). The strategy was to minimize the HSP distance of the candidate molecules from furfural. To reflect practical considerations, the HSP distance between toluene and furfural (13.9) was set as the tolerance value, given that toluene is a viable solvent for furfural extraction.

For each configuration of fixed-building blocks and kmax (maximum number of fragments added in the fixed building block structure, varied from 4 to 8), a different input XML file was created. These files included the target properties (Table 1) and the variable fragments (-CH2-, -CH--, --C--, CH2=, -CH=, --C=, CH=-, -C=-, NH2-, -NH-, -N--, NH=, N=-, OH-, -O-, O=).

Properties calculation

Using group contribution methods for property estimation, the tool is able to predict the melting point, boiling point and log(kow) with models developed by Hukkerikar *et al.* [16]; HSP distance with molecular breaking by Hiroshi Yamamoto [17] and flash point with a model created by Catoire, Paulmier and Naudet [18].

Performance criteria for molecule evaluation

The performance of each candidate molecule is calculated from a performance function (Eq. 1), and it considers deviation of the predicted property values from their target values (Eq. 2).

$$Perf\ (MG_i) = \frac{\sum_{p=1}^{np} w_p.F_p.(MG_i)}{\sum_{p=1}^{np} w_p} \quad (1)$$

$$F_p(MG_i) = exp.\left[\ln(val).\left(\frac{P-x}{Tol}\right)^2\right] \quad (2)$$

In the equations, Mg_i is the molecular graph, $Perf$ is the calculated performance, n_p is the number of target properties, F_p is the performance function, w_p is the weight of property p, Tol is the tolerance set, x is the target value for property and P is the property value.

Optimization

The problem solved by IBSS® is a multi-objective optimization since multiple properties are considered in the performance equation. A genetic algorithm is used, which modifies the molecular graphs through mutation, crossover, deletion, and insertion. Newly generated molecules have their performance calculated and are ranked accordingly. The molecules with the best performance are selected for the next generation. The parameters of the genetic algorithm, presented in Table 2, were used in this study.

Results Analysis

Table 1: Target properties setup.

Property	Weight	Unit	Target	Tolerance	Cor. Perf.[1]	Justification	SD[2]	Ref.
HSP Distance	3	$MPa^{1/2}$	0	13.9	0.9	Good solubility	-	[17]
Melting Point	1	°C	< 10	10	0.8	Liquid at usage temperature	21,45 K	[16]
Boiling Point	1	°C	>250	3	0.8	Liquid at usage temperature	7,67 K	[16]
Flash Point	1	°C	>61	3	0.8	Non flammable	-	[18]
Log (Kow)	1	-	<3	0.8	0.8	Low potential of bioaccumulation	0,61	[16]

[1]Cor. Perf. = Corresponding Performance; [2]SD = Standard Deviation

The third stage involved an evaluation of the results to identify the most promising candidates. A total of 16 thousand molecules were generated in the final generations and more than 3 million properties were calculated throughout the study. The main results are presented below.

Table 2: Parameters of the genetic algorithm.

Property	Value
Elitism	5
Population size	100
Number of generations	200
Crossover probability	65%
Mutation probability	15%
Insertion probability	10%
Deletion probability	10%

RESULTS

The computational exploration conducted using the IBSS® CAMD tool enabled the evaluation of millions of potential molecules, leading to the identification of promising alternatives to toluene for LLE of furfural. The best molecules for each fixed building block derived from orange and coffee wastes were ranked based on their performance and the outcome is presented in Table 3.

A set of alternative methods was tested for predicting properties: log(KoW), melting point and boiling points were predicted using a method by Marrero&Gani [19]; and HSP Distance was calculated using a method by Hukkerikar *et al.* [16]. The average variation in the calculated performances was 13%. This highlights the robustness of the property predictions, despite the inherent limitations of group contribution methods. Further analysis could explore how these variations impact the selection of candidate molecules in practical applications.

The results indicate that derivatives of glycerol, limonene, and furfural performed best. Nevertheless, 13 configurations exhibited high performances (above 0.9800), a promising outcome that suggests multiple alternative solvents worth exploring.

Toluene was evaluated in IBSS®, in evaluation mode, applying the same target properties outlined in Table 1. The calculated performance was 0.6724, lower than the performance of the generated molecules presented in Table 3. This low value is primally driven by the low flash point (8°C), low boiling point (116°C) and high HSP distance (13.9 $MPa^{1/2}$), indicating that the bio-based molecules generated have strong potential to serve as superior solvents for furfural extraction.

Analyzing the results in Table 3, it is not possible to establish a direct connection between number of radicals and performance, but it is noticed that for limonene, hydroxypropionic and glutamic acid, an increase in the number of functionalization points was positive to the performance and the opposite behavior was identified for glycerol, hydroxymethylfurfural, methylfurfural, geranyl and succinic acid.

Also, it is observed that the biggest molecules, in general, did not score high performances. The average of non-hydrogen atoms for the molecules with performances greater than 0,9 was 14, while for the molecules with performances less than 0,9, it was 26.

Three of the best 26 molecules presented in Table 3 were selected and their properties are detailed in Table 4. The values of the properties are highlighted in green if they meet the target value, or in blue if they don't meet the exact target but fall within the acceptable range defined by the target plus the tolerance.

Table 3: Best molecules generated for each building block.

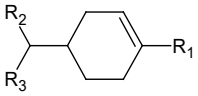
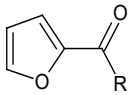
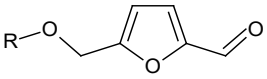
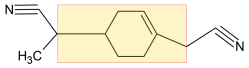
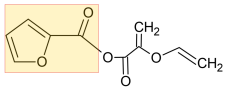
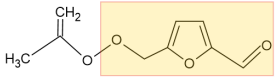
Fixed Building Block (FP[2], Kmax)	B. P.[1]	Fixed Building Block (FP[2], Kmax)	B.P.
Glycerol (1, 8)	0.9986	Geranyl (2, 4)	0.9817
Glycerol (2, 4)	0.9955	Methylfurfural (2, 4)	0.9810
Limonene (3, 4)	0.9937	Glycerol (5, 6)	0.8571
Furfural (1, 8)	0.9931	Glutamic Acid (2, 5)	0.8571
Furoate (1, 7)	0.9931	HMF[4] (3, 5)	0.8571
Limonene (2, 4)	0.9930	Glutamic Acid (1, 7)	0.8568
HP[3] (3, 4)	0.9925	Succinic Acid (1, 8)	0.8568
HMF[4] (1, 4)	0.9922	Glycerol (3, 7)	0.8567
Levulinic Acid (1, 5)	0.9914	Succinic Acid (3, 5)	0.8561
HP[3] (2, 4)	0.9913	Glucaric Acid (6, 7)	0.8552
Succinic Acid (2, 6)	0.9904	Oleic Acid (2, 6)	0.8550
Glutamic Acid (4, 4)	0.9883	Glycerol (6, 6)	0.8548
Limonene (1, 7)	0.9876	Succinic Acid (4, 7)	0.8541
Methylfurfural (1, 5)	0.9870	Oleic Acid (1, 7)	0.8510
HMF[4] (2, 4)	0.9848	HP[3] (1, 7)	0.8496
Geranyl (1, 8)	0.9843	-	-

[1]BP = best performance
[2]FP = functionalization points
[3]HP = hydroxypropionic
[4]HMF = hydroxymethylfurfural

In all three cases, the Hansen Solubility Parameter (HSP) distances were significantly smaller than the toluene-furfural distance of 13.9. Additionally, their melting and boiling points suggest that these solvents would remain liquid under ambient conditions and based on their flash point they are classified as non-flammable. Furthermore, their low log(Kow) values indicate minimal bioaccumulation concerns, supporting their suitability as sustainable solvent alternatives.

As shown in Figure 3, the maximum performance of the best molecule, derived from glycerol with one functionalization point, improved as generations progressed. This trend demonstrates the effectiveness of the genetic algorithm in identifying better-performing individuals across the tested generations. Furthermore, the results reveal that as kmax increases, the maximum

Table 4: Comparative analysis of molecules derived from limonene, furfural and hydroxymethylfurfural.

Property	Limonene Derivative	Furfural Derivative	Hydroxymethylfurfural Derivative
Fixed building block structure			
New molecule structure			
Fixed building block	Limonene	Furfural	Hydroxymethylfurfural
Number of FP*	3	1	1
Kmax**	4	8	4
Performance	0.9937	0.9931	0.9922
Melting point (°C)	11.95	10.09	10.06
Boiling point (°C)	304.99	286.48	264.69
Flash point (°C)	152.96	148.42	126.33
HSP distance $(MPa)^{1/2}$	4.69	5.46	5.79
Log (kow)	1.92	0.66	1.06

*FP: functionalization point
**Kmax: maximum number of fragments allowed to be added in the fixed building block

performance also improves in this case, highlighting the algorithm's enhanced flexibility to create diverse molecular structures with higher potential.

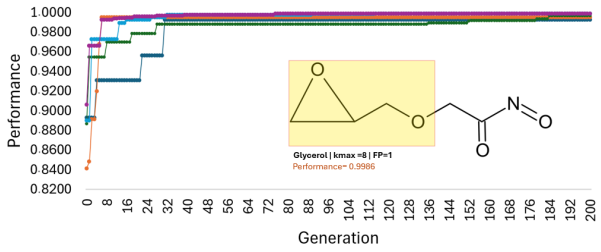


Figure 3: Performance evolution across generations for glycerol derivative with 1 functionalization point and representation of the best molecule generated.

From the list in Table 3, some molecules were selected and drawn based on the molecular graph provided as an output of IBSS®. These molecules are depicted in Figure 4. The chemical structures are diverse, and it is important to emphasize that further investigations are needed into process design, stability, and the viability of the proposed molecules. Nevertheless, this study highlights the potential of utilizing residues from high-demand fruits to design environmentally friendly solvents with superior performance characteristics.

CONCLUSION AND NEXT STEPS

The results demonstrate that the IBSS® CAMD tool is an effective approach for designing bio-based solvents with specific properties to replace toluene in LLE of furfural from aqueous solutions. Through multi-objective optimization, alternative solvents were efficiently ranked based on a global performance value. For all the building blocks studied, molecules with enhanced desired properties compared to toluene were obtained, highlighting the derivatives of glycerol, limonene, and furfural itself. While this early-stage methodology shows significant promise, further studies are required to assess the stability and feasibility of synthesizing these molecules.

Nonetheless, those findings demonstrate a valuable pathway for developing alternative solvents derived from orange and coffee waste, offering a sustainable solution to reduce reliance on fossil-based solvents.

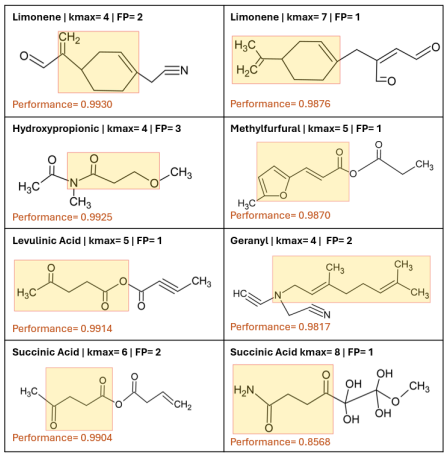


Figure 4: Molecular representation of molecules generated by IBSS®.

ACKNOWLEDGEMENTS

This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brazil. Process Number 2024/03274-9 and 2022/02809-0.

REFERENCES

1. Brasil. Systematic survey of agricultural production. Instituto Brasileiro de Geografia e Estatística (IBGE). https://www.ibge.gov.br/en/statistics/economic/agriculture-forestry-and-fishing/17174-systematic-survey-of-agricultural-production.html?=&t=resultados (Accessed Dec. 19, 2024).
2. Capanoglu E, Nemli E, Tomas-Barberan F. Novel approaches in the valorization of agricultural wastes and their applications. *J Agric Food Chem* 70:6787-6804 (2022). https://doi.org/10.1021/acs.jafc.1c07104
3. Brasil. Coffee production explained. Instituto Brasileiro de Geografia e Estatística (IBGE). https://www.ibge.gov.br/explica/producao-agropecuaria/cafe/br (Accessed Jan. 15, 2025)
4. Lee YG, Cho E, Maskey S *et al.* Value-added products from coffee waste: a review. *Molecules* 28:3562 (2023). https://doi.org/10.3390/molecules28083562
5. Aristizábal-Marulanda V, Chacón-Perez Y, Alzate CA. The biorefinery concept for the industrial valorization of coffee processing by-products. *Handbook Coffee Process By-Prod* 63-92 (2017). https://doi.org/10.1016/b978-0-12-811290-8.00003-7
6. United States Department of Agriculture. Foreign Agricultural Service. Brazil citrus semi-annual. https://fas.usda.gov/data/brazil-citrus-semi-annual-7 (Accessed Oct. 19, 2024).
7. Suzuki LL, Cunha I, Teixeira C *et al.* The modeling and simulation of waste recovery processes applied to the orange juice industry. *Sustainability* 15:15225 (2023). https://doi.org/10.3390/su152115225
8. National Renewable Energy Laboratory (NREL). Chemicals from Biomass: Market Assessment of Bioproducts with Near-Term Potential. https://www.nrel.gov/docs/fy16osti/65509.pdf (Accessed Dec. 13, 2024).
9. Sai MSN, De D, Satyavathi B. Sustainable production and purification of furfural from waste agricultural residue: an insight into integrated biorefinery. *J Clean Prod* 327:129467 (2021). https://doi.org/10.1016/j.jclepro.2021.129467
10. Gerbaud V. Computer Aided Molecular Design for bio-sourced molecules. The IBSS Tool. *Journée d'échange Formul Agents Subst Biocontrol* (2023).
11. Heintz J, Touche I, dos Santos MT, Gerbaud V. An integrated framework for product formulation by computer aided mixture design. *Computer Aided Chemical Engineering* 30:702-706 (2012). https://doi.org/10.1016/B978-0-444-59519-5.50141-6
12. Rodriguez-Donis I, Gerbaud V *et al.* Computer-aided product design of alternative solvents based on phase equilibrium synergism in mixtures. *Comptes Rendus Chim* 21:606-621 (2018). https://doi.org/10.1016/j.crci.2018.04.005
13. Wang J, Rodriguez-Donis I, Thiebaud-Roux S *et al.* Selection of green solvents for organic photovoltaics by reverse engineering. *Mol Syst Des Eng* 7:182-195 (2022). https://doi.org/10.1039/D1ME00151E
14. Nehmeh M, Rodriguez-Donis I, Cavaco-Soares A *et al.* Bio-refinery of oilseeds: oil extraction, secondary metabolites separation towards protein meal valorisation—A review. *Processes* 10:841 (2022). https://doi.org/10.3390/pr10050841
15. Heintz J, Belaud JP, Pandya N *et al.* Computer-aided product design tool for sustainable product development. *Comput Chem Eng* 71:362-376 (2014). https://doi.org/10.1016/j.compchemeng.2014.09.009
16. Hukkerikar AS, Sarup B, Sin G *et al.* Group-contribution+ (GC+) based estimation of properties of pure components: improved property estimation and uncertainty analysis. *Fluid Phase Equilib* 321:25-43 (2012). https://doi.org/10.1016/j.fluid.2012.02.010
17. Abbott S, Hansen CM, Yamamoto H. Hansen solubility parameters in practice – complete with software, data, and examples. 5th ed. (2015). ISBN: 9780955122026. Available from: www.hansen-solubility.com.
18. Catoire L, Paulmier S, Naudet V. Estimation of closed cup flash points of combustible solvent blends. *J Phys Chem Ref Data* 35:9-14 (2006). https://doi.org/10.1063/1.1928236
19. Marrero J, Gani R. Group-contribution based estimation of pure component properties. Fluid Phase Equilib 183:183-208 (2001). https://doi.org/10.1016/S0378-3812(01)00431-9