Using of Multidimensional Data Analysis and Neural Network Approaches.

Mourad Korichi[1,2], Vincent Gerbaud[2], Pascal Floquet[2], A.-Hassan Meniai[3], Saci Nacef[4] and Xavier Joulia[2]

[1] LVPRS, Ouargla University, Department of Chemical Engineering, LVPRS, BP 511 Ouargla ALGERIA, M.Korichi@arn.dz

[2] Laboratoire de Génie Chimique UMR 5503, 5 rue Paulin Talabot, BP 1301, 31106 Toulouse Cedex 01, FRANCE, Vincent.Gerbaud@ensiacet.fr

[3] LIPE, Constantine University, Constantine 25000 and [4] LGC, Sétif University, Sétif 19000, ALGERIA

I. INTRODUCTION

- **Odorant compounds** are found in a wide variety of products ranging from foods, perfumes, health care products and medicines. Either combined or alone, flavor and fragrance compounds are used to induce consumers to associate favorable impressions with a given product. In some cases, products have one predominant component which provides the characteristic odor.
- In most cases, products containing odors include a complex mixture of fragrant compounds. Some of them are classified within **REACH**, the European Community document regulating the use of chemicals in terms of environment and toxicity.
- **Structure – Odour relationships (SOR)** are very important for the synthesis of new odorant molecules. This relation is difficult to model due to the subjectivity of the odor quantity and quality (see Table 2). **Olfaction phenomenon is not yet completely understood and odor measurements are often inaccurate** (Amboni *et al.*, 2000). Research has been oriented to the use of structural, topological, geometrical, electronic, and physicochemical parameters as descriptors, to generate odor predictive equations.

II. GOAL

we aim to use molecular descriptors as an alternative approach in the prediction of molecule's odour by the mean of classification and regression techniques.

III. METHODOLOGY

- Our input data set contains 121 molecules of balsamic odour splitted in 5 sub-notes of typical odors: anise, balsam, honey, vanilla and sweet (Aldrich Flavors and Fragrances catalog, 2005) (see Table 1).

Table 1. Input data set of molecular structure

Odor	Number of compounds	Arbitrary continuous Odor codification	Arbitrary discontinuous Odor codification
Anise	10	0 to 0.15	0.15
Balsam	18	0.25 to 0.35	0.35
Honey	21	0.45 to 0.55	0.55
Vanilla	15	0.60 to 0.75	0.75
Sweet	58	0.85 to 0.95	0.95

- The dragon software (TALETE, 2005) is used to calculate the molecular descriptors of the input data set molecules.
- Figure 1 summarizes the proposed methodology. According to figure 1, two cases are explored (All descriptors are calculated from the 2D molecular representation.):
  - ✓ ***Molecular descriptors calculation***
    - o first, 34 simple descriptors (molecular weight, aromatic ratio, number of benzene-like rings, connectivity indices, Kier symmetry index and n-path Kier alpha-modified shape index) are calculated.
    - o In the second case, 582 molecular descriptors are considered: constitutional descriptors, topological indices, walk and path counts, connectivity and information indices, 2D autocorrelations, edge adjacency indices, burden eigenvalues and topological charge indices.
  - ✓ ***Complete Correlation Analysis***

    select a subset of linearly independent descriptors. Descriptor dependency is evaluated using the Dragon software by setting a predefined value R_{max} (In this work, R_{max} = 0.97) below which descriptors are considered linearly independent.
  - ✓ ***Principal component analysis***

    It enables us to reduce the molecular descriptors dimensionality by the identification of principal components that can be used in the structure - odor relationship. All eigenvalues > 1 are retained to describe the principal axes.

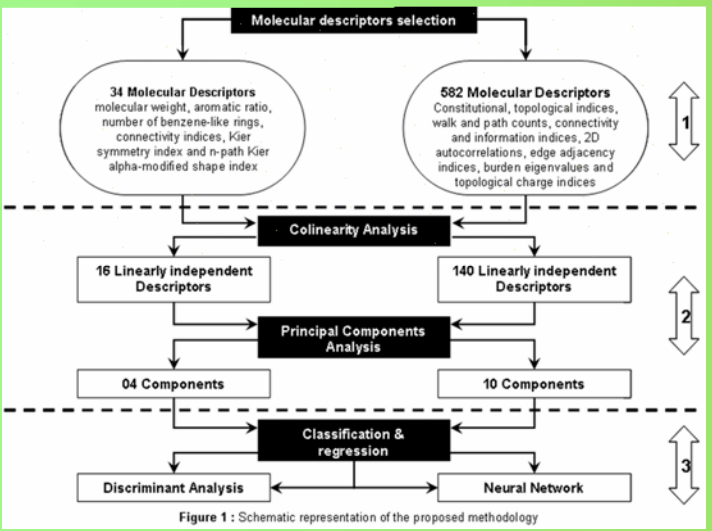
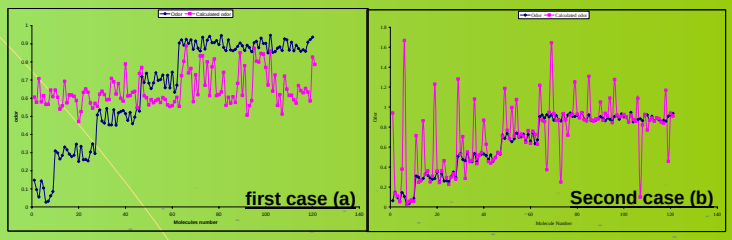


Figure 1 : Schematic representation of the proposed methodology

IV. RESULTS AND DISCUSSIONS

- **Artificial Neural Networks Approach (**Arbitrary continuous odor codification**)**

  In the figure below, results are presented as the variation of the odor quality versus the molecule id code.



  - ✓ **In the first case (a)**, the ANN does not converge as shown by the similarity of the response for all molecules despite their initial differences.
  - ✓ **In the second case (b)**, all errors are found among the validation set whereas the training set is well represented.

  This clearly shows the non predictive capacity of the ANN approach.

- **Discriminant Analysis Approach (**Arbitrary discontinuous odor codification**)**
  - ✓ 69.4% and 83.4% of molecules are well discriminated in the two cases respectively (see table 2 & 3).
  - ✓ Case 2 has more odor discriminant, because it incorporates more descriptors.
  - ✓ Among the molecules that are not discriminated, the two molecules from anise, classified in the vanilla group bear similar molecular structure with vanilla type molecules, which have three oxygen atoms, high Kier symmetry index and n-path Kier alpha-modified shape index.
  - ✓ For case 2, in balsam and honey sub-note odors, the molecule wrongly classified is considered differently depending on the referential nomenclature.
  - ✓ In the sweet sub-note, fourteen molecules are distributed into other sub-notes. The low discrimination of the sweet odor may be attributed to the subjectivity of this sub-note, unlike vanilla or anise. Indeed, sweet is not considered as a typical odor type in the reputed referential chart "the field of odors" of Jaubert et al.

Table 2. Discriminant analysis based on the first PCA study.

Groups	Predicted groups						Correctly classified
	01	02	03	04	05	molecules	
Anise (01)	8	0	0	2	0	10	0.800
Balsam (02)	1	14	1	1	1	18	0.778
Honey (03)	0	3	15	0	3	21	0.714
Vanilla (04)	0	2	0	12	0	14	0.857
Sweet (05)	3	8	8	4	35	58	0.603

Table 3. Discriminant analysis based on the second PCA study

Groups	Predicted groups						Correctly classified
	01	02	03	04	05	molecules	
Anise (01)	8	0	0	2	0	10	0.800
Balsam (02)	0	17	0	0	1	18	0.944
Honey (03)	0	0	20	0	1	21	0.952
Vanilla (04)	0	1	0	13	0	14	0.929
Sweet (05)	3	2	5	4	44	58	0.759

Table 4. The subjectivity of odors

Selon la base de données d'Aldrich Inc.	Selon le référentiel Champs des Odeurs®
❖ **Balsamic**	
✓ Anise	❖ **Balsamic**
✓ Balsam	❖ **Anise**
✓ Honey	❖ **Honey**
✓ Vanilla	❖ **Vanilla**
✓ Sweet	

V. CONCLUSION AND PERSPECTIVES

- In this work we present different ways to estimate and discriminate odors.
- Discriminant analysis results using only 2D molecular representation are encouraging. Further work using 3D representation molecular descriptors may improve the results.
- The neural network satisfactorily correlates the molecules with their assigned odor, based on sufficiently numerous and diverse molecular descriptors. But it is unable to predict balsamic odor and its sub-notes. Compared with literature, successful results in ANN approach are due to the well known families of odor.
- The heterogeneous nature of the molecules assigned to balsamic odor and the absence of evident structure – odor relationship, forces us to request a continuous discrimination between sub-notes.