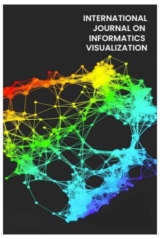
INTERNATIONAL JOURNAL
ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv



A Classification Algorithm Inspired by the Chromatographic Separation Mechanism Dedicated to the Classification of Variable-length and Multi-class Vectors

Mariusz Święcicki [a,*]

[a] *Cracow University of Technology, Warszawska 24, Cracow, 31-155, Poland*
Corresponding author: [*]mariusz.swiecicki@pk.edu.pl

Abstract*—Nowadays, one of the critical problems related to data mining is the processing of large data sets. This article presents an algorithm that may apply to the issues associated with classifying large-volume data sets. The motivation behind defining this type of algorithm was that the methods used to process this data type are subject to several significant limitations. The first considerable limitation of using classical classification methods is ensuring a constant data size. The second type of constraint is related to the data dimension. The last limitation in using classic classification algorithms is associated with the situation in which a given input vector may contain data belonging to many classes simultaneously, in which case we are talking about so-called multi-class vectors. The presented algorithm is inspired by the method of chromatographic separation of chemical substances. This method is widely and successfully used in analytical chemistry. As we know, in the case of chromatographic separation, we are dealing with a similar class of problems that occur when processing large data sets, firstly: the molecules of a chemical substance have a different number of molecules - i.e., they have different lengths, which corresponds to the situation that occurs when processing large data sets. In this work, a classification algorithm inspired by the mechanism of resolution chromatography is presented. The article presents the results of calculations for sample data sets. It discusses issues related to the properties of the defined algorithm, which concern the algorithm training process and the classification of single-class and multi-class data.

***Keywords*— Natural computing algorithms; chromatographic separation; chromatographic data classification; signal processing; data mining; big data sets**

Manuscript received 7 Nov. 2023; revised 17 Jan. 2024; accepted 14 Feb. 2024. Date of publication 31 Mar. 2024.

I. Introduction

This article presents an algorithm that may apply to issues related to the classification of large-volume data sets. The motivation behind defining this type of algorithm was that the methods used to process this data type are subject to several significant limitations. The first considerable limitation in using classical classification methods is the need to ensure a constant size of data - vectors that will be subject to the classification process [1], [2]. The second type of constraint is related to the data dimension. When we use classical methods to classify large vectors, we always have to reduce the dimension of the input vectors by using selected methods of mathematical statistics. Another limitation of the algorithms currently used is that the classified data must be homogeneous, i.e., only one type of data can exist. If images are classified, data that are not images and whose source of data is another phenomenon that is somehow related to the classified photos cannot be classified as input data simultaneously. Finally, the last type of limitation that occurs in the use of classic classification algorithms is related to the situation that a given input vector may contain data belonging to many classes at the same time, in which case, in this article, we are talking about so-called multi-class vectors [2].

The presented algorithm attempts to solve the problems defined above. The chromatographic separation of substances inspired the algorithm used in analytical chemistry [3]. The article's first chapter will present the chromatographic data separation principle, which is the basis of the defined algorithm. The second chapter will present an algorithm for data classification inspired by the principles of chromatographic separation of substances. The following parts of the article will show the results of calculations and classification, for example, data se

A. Principle of Chromatographic Separation

In general, separation is a process in which a mixture of chemical compounds is divided into at least two fractions of different compositions [4], [5], [6]. From a chemical point of view, the purpose of the substance separation process is to increase the concentration of one component of the initial mixture relative to the other components of the initial mixture. Separation is achieved by using physical methods as well as chemical reactions [4], [6], [7], [8], [9], [10], [11].

Based on the assumptions given above, it can be concluded that the chromatographic system consists of the following elements: a stationary phase called the stationary phase, a mobile phase, i.e., a mixture of separated substances [5], [6], [11], [12], [13], [14], [15], [16]. The elements mentioned are present in every chromatographic system. The condition that must be met to separate substances is the movement of one phase relative to the other.

Chromatography is a physicochemical separation method in which the separated components are divided between two phases, stationary (stationary phase) and mobile (mobile phase), which move in a specific direction. The different division of the substance of the mixture between both phases causes different migration rates and separation of the components [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [16], [17], [18], [19]. For example, the mobile phase moves inside the column, while the stationary phase is deposited on the inner walls of the column. Separated substances, i.e., chemical compounds with a greater affinity for the stationary phase, are selectively retained by it and move along the column much slower. However, chemical compounds with lower affinity for the stationary phase move along the column faster and thus leave the column, i.e., elute from the column, first. The equilibrium of partitioning between phases is dynamic, i.e., molecules of a substance constantly move from the mobile phase to the stationary phase and back.

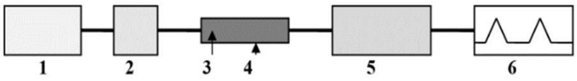


Fig. 1 Block diagram of the chromatograph [15,16]

Figure 1 shows the block diagram of the chromatography system. The stationary phase (3) is located in column (4), through which the mobile phase flows and is pumped by the pump (1). The chromatographed mixture is introduced into the column (4) using a dispenser (2). The mix of substances separates in the column, and the first component that travels fastest flows into the detector (5). The detector is sensitive to the change in concentration in the mobile phase and when the substance appears in the detector (concentration change), which is recorded by the graphic recorder (6). The recorder records all changes in the concentration of successively appearing, separated substances in the form of a series of peaks, which we call a chromatogram.

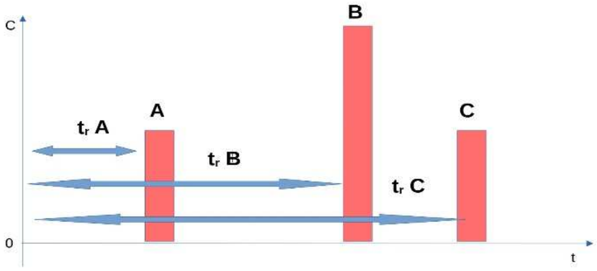


Fig. 2 The process of creating a chromatogram

The fundamental quantity we use to separate substances during substance separation is the retention time tr, i.e., the time that has elapsed from introducing the substance into the system until the peak maximum appears in the detector. In Figure 2, the arrows show the retention times for substances A, B, and C [6], [13], [14], [15], [20].

B. Assumption of the Defined Chromatographic Data Separation Algorithm

The chromatographic data separation algorithm is based on the basic paradigm that the processed data string is a complex chemical molecule with a chain-linear structure. This means that each data vector or set of vectors will be processed by the chromatographic algorithm by the rules that apply in the real chromatographic system.

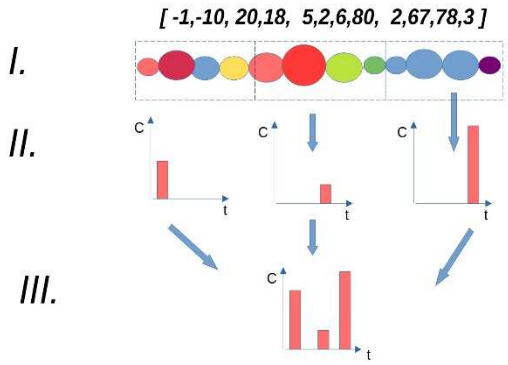


Fig. 3 Operations of the chromatographic data classification algorithm

The general principle of operation of the chromatographic data separation algorithm will be to treat the data vector as a mixture of chemical compounds. For each "chemical" compound, the relationship between the concentration of a given substance at the output of the chromatographic system is calculated. In other words, it will involve calculating the spectrum, as shown in the figure. In the first phase, we treat the vector of numbers as a polyatomic molecule with a linear structure. In the next phase, the molecule is divided into smaller molecules. In the last phase of the algorithm, the chromatographic column processes each newly created molecule, i.e., the retention time is calculated. A chromatogram is created due to these operations, i.e., a graph describing the concentration of a given type of molecule as a function of time at the output of the "chromatographic" column. This relationship, i.e., the chromatogram, is later called the spectrum of a given starting substance.

II. Material and Method

The chromatographic data separation algorithm consists of the following sequence of operations, which are inspired by the functioning of a real chromatographic system:

- Mixture creation phase for a given vector
- Retention time calculation phase
- Chromatogram creation phase
- Spectrum analysis phase

A. The Phase of Creating a Mixture for a Given Vector and the Phase of Calculating the Retention Time

In the first phase of this algorithm, a set of vectors W consisting of any number of vectors of any length is transformed into a set of mixtures of substances by dividing the fragmentation into smaller vectors of the same size. The fragmentation of the vector takes place so that for each element from the set *W*, a mixture of substances is created that corresponds to this element of the set *W*.

```
Input data
$W=\{w_1,w_2,w_3....w_N\}$ – a set of data vectors that will be processed

Output data
$CH=\{ch_1, ch_2, ch_3....ch_N\}$ – a set of chromatograms, where each
element of this set represents a chromatographic spectrum
corresponding to a given aspect of the set W

$MS_{i..M}=[];$
$W=\{w_1, w_2, ......w_M\}$

Foreach  w ϵ W

For a given $w_i$ data vector, create a mixture of substances—this will
fragment the vector into sub-vectors of constant length.
$MS_i=\{s_1, s_2, ......s_{M(i)}\}$
$MS_i$ -a set of substances is created by dividing a vector into sub-
vectors according to the adopted principle of division,

$ms_i$ - the elements of this set is the set of substances resulting from the
division of the vector $w_{i;\ this}$ means that the set will contain individual
substances s that are not subject to further subdivision
$ms_{M(i)}:=\{s_1, s_2, ......s_{M(i)}\}$
 a substance that was created by splitting the wi vector. Wi.

    Foreach  s ϵ $ms_i$

        Calculate Retention Time i $t_r$ – the residence time of the
        substance in the stationary phase.

    End
end
```

Algorithm. 1 Algorithm transforming a set of vectors into a set of chromatograms

As shown in algorithm 1, the set of created mixtures of substances is fed to the input of a "virtual chromatographic column," in which a given substance migrates between the stationary phase and the mobile phase. The value of the retention time t_r depends on the affinity of the stationary phase for a given substance, which is an essential value in the classification process.

B. Chromatogram Formation Phase

The next stage of the presented algorithm is creating a chromatogram for a given mixture of substances corresponding to the w_i element. The chromatogram is created by registering individual substances at the output of the chromatographic column. The moment a given substance will leave the chromatographic column depends on the retention time t_r. The purpose of the detector is to count the molecules of substances leaving the chromatographic column at a given moment in time.

```
For a given set of substances $MS_i$ that a chromatographic column
has processed, i.e., they have a calculated retention time $t_r$.

$ch_i=[];$

 Foreach  s ϵ $MS_i$

$peak_i[$ s.Tr $]:=peak_i[$ s.Tr $]+1$

end
```

Algorithm. 2 Algorithm for creating a ch_i chromatogram for a mixture belonging to the w_i vector

Algorithm 2 above presents the process of detecting a substance that leaves the chromatographic column. This algorithm creates a ch_i chromatogram for the element w_i.

C. Learning Mechanism

Using the algorithms presented above, we can create a set of chromatograms of known substance mixtures corresponding to a given vector from the set of input vectors.

D. Spectrum Analysis Phase

The last stage of recognizing substances that the chromatographic system has processed is classifying the output chromatographic spectrum and assigning it to the spectra of known substances.

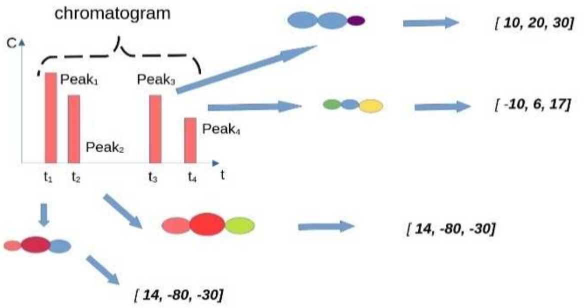


Fig. 4 Structure of the chromatogram - the spectrum corresponding to the w_i vector

The chromatogram of the tested mixture of substances describes the concentration of individual compounds that constitute the composition of the tested substance, which were separated as a result of the chromatography process, similar to the presented algorithm. As shown in the figure above, as a result of the chromatography process, we obtain a chromatogram that contains many peaks corresponding to the concentration of "substances" created as a result of the operation of algorithm 1. The classification algorithm task will assign the chx chromatogram to the chromatograms of known vectors, using the matching criterion, which is the retention time. Peak.

The input data for algorithm 3 is the ch_x chromatogram, which will be compared with the elements of the set of *CH* chromatograms corresponding to individual classes. The presented algorithm calculates a match to the chx chromatogram for each chi element belonging to the set of CH chromatograms. Matching the ch_x chromatogram to the chi chromatogram consists of matching all peaks belonging to the chx chromatogram for which the difference between the retention time of the jth peak of the chi chromatogram is smaller than the set value, i.e., *eps*. Two values are calculated for those peaks that meet the above criterion.

```
0    **Input data**
     $ch_a=\{peak_1,peak_2,peak_3....peak_N\}$ – a chromatogram consisting of N peaks
     $CH=\{ch_1,ch_2,ch_3....ch_l\}$
     $ch_i=\{peak_1,peak_2,peak_3....peak_M\}$ –a chromatogram consisting of M peaks

     **Output data**
     D - Distance is a value that determines the level of similarity between the $ch_a$, and $ch_b$ chromatograms
     NoClass – class number
1    NoClass:=0; MinDist:=∞
2    Foreach $ch_i$ є CH
3        Foreach $peak^i_j$ є $ch_i$
4            $P:=\{ peak^i_j | abs(t_r{}^i_j - t^x) < eps \}$
5        $f_i:=f_i+sum( peak^i_j)$
6        $f_x:=f_x+sum( P)$
7        end
8        $f_i:=f_i/sum( peak^i_j)$
9        $f_x:=f_b/sum( peak^x_i)$
10       $d_i:=sqrt( (1-f_i)^2 + (1-f_x)^2 )$
11       D:=D □ $d_i$
12   end
13   NoClass:={i | min ( $D_{1..N}$) = $D_i$ }
```

Algorithm. 3 An algorithm for classifying a vector wx using its chx chromatogram

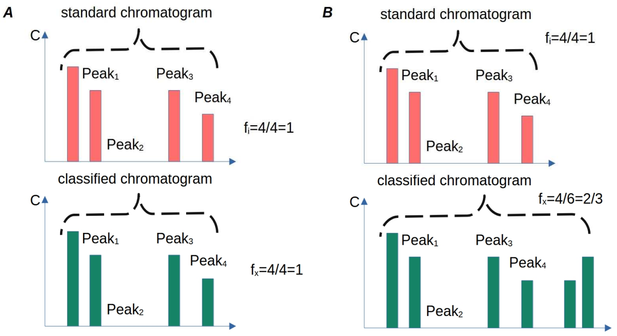


Fig. 5 Peak matching process

Value the phi value (line 8), which specifies the fraction of fitted peaks from the chi chromatogram to the *ch_x* chromatogram. However, in line 9, the second f_x is also calculated, which contains information about what part of the peaks from the classified ch_x chromatogram was assigned to the peaks from the i-th chromatogram. Line 10 calculates the matching value between the number of matched peaks for the tested chromatogram and the number of peaks matched in the reference chromatogram. As follows from the presented algorithm, the selected chromatogram for which the formula specified on line 10 takes the minimum value means that it will be the chromatogram for which the most significant number of peaks belonging to the tested chromatogram were matched and, at the same time, the most considerable number of peaks were matched in the reference chromatogram.

The calculation of the f_x and f_i values is shown in the figure above. Figure 5 shows that if all peaks from the standard chromatogram I and the classified chromatogram are matched, then the f_x and f_i values are equal to one, as shown in the figure in part A. However, part B shows when the classified chromatogram contains more peaks than the reference. Then, the f_x and f_i values determine the fraction of peaks fitted in the classified chromatogram, and the f_i value specifies the number of peaks used from the reference chromatogram. The situation shown in the figure will occur during the classification of multi-class vectors.

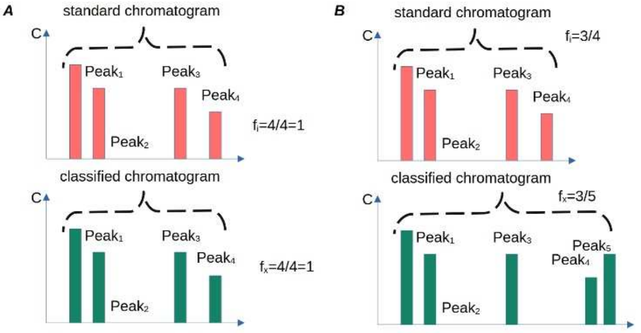


Fig. 6 Peak matching process

The following Figure 6, part B, illustrates the peak matching process when the matching of the classified chromatogram to the reference chromatogram is incomplete. This situation occurs when the chromatogram is classified as a result of processing an input vector whose elements have been distorted compared to the standard vector. The algorithm assigns the chromatogram for which the values in line 8 and line 9 reach the maximum value to the classified chromatogram.

E. Problems of Selecting the Stationary Phase

There are two significant problems when performing calculations using the algorithm presented above. The first problem, already indicated in the previous chapter, is related to the stationary phase selection so that the chromatograms of vectors belonging to different classes are characterized by different retention times. The second problem, in a sense, is a derivative of the first problem and is related to the fact that the chromatograms that are created in the process are complex, i.e., they contain a large number of peaks, which makes the classification process difficult by the presented algorithm classifying chromatograms [5], [6], [11], [13], [14], [15], [18], [20], [21], [22], [23].

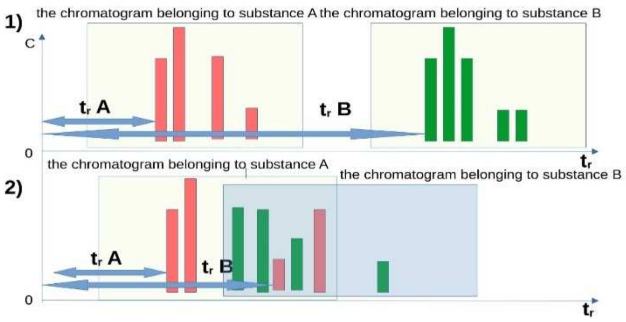

1)
the chromatogram belonging to substance A
the chromatogram belonging to substance B
C
tᵣ A
tᵣ B
0
tᵣ
2)
the chromatogram belonging to substance A
the chromatogram belonging to substance B
tᵣ A
tᵣ B
0
tᵣ

Fig. 7 The phenomenon of overlapping retention times

At this point, an analysis of the algorithm's functioning will be carried out, considering the problem of selecting the stationary phase for a given set of input data vectors; for this purpose, the following notations will be introduced. Let us assume that the stationary phase FS is an m-element vector, as shown in equation (1). At the same time, the substance vector that was created as a result of the algorithm in the fragmentation process as a result of the operation of the first algorithm 1 is also an array with dimensions NxM presented in equation (2).

$$FS = (fs_1, fs_2, fs_3 \dots fs_M) \tag{1}$$

$$S = \begin{bmatrix} s_{1,1}, s_{1,2}, s_{1,3} \dots \dots s_{1,M} \\ s_{2,1}, s_{2,2}, s_{3,1} \dots \dots s_{3,M} \\ \dots \dots \dots \\ s_{N,1}, s_{N,2}, s_{N,1} \dots \dots s_{N,M} \end{bmatrix} \tag{2}$$

As we know, a chromatogram is made up of peaks, and a single peak is a pair of numbers, the first of which is the retention time tr and the second is the concentration of substance *C*, formula (3)

$$peak_i = (tr_i, C) \tag{3}$$

The retention time can be calculated using the Ftr function, which calculates the retention time value for a given substance and the vector describing the stationary phase (4)

$$tr_i = F_{tr}(S_{i,1\dots M}, FS) \tag{4}$$

When calculating the retention time, the function calculates the retention time for a given substance, taking into account the structure, i.e., the values of the stationary phase. For further consideration, it can be assumed that the function calculating the retention time is expressed by formula (5).

$$tr_i = F_{tr}(S_{i,1\dots M}, FS) = \sum_{k=1}^{M} (s_{i,k} \cdot fs_k) \tag{5}$$

As the presented formula shows, the scalar product of two vectors is calculated. The more similar the vectors are to each other, the greater the value of the calculated product and the greater the retention time for a given substance.

The description of the algorithm and the drawing above show that the correctness of classification is significantly influenced by the distribution of peaks in the chromatogram of the reference substance and the chromatogram of the classified substance. The optimal situation occurs when the distances between individual chromatograms are large or, in other words, the peaks of individual substances do not overlap. The formula describing the distance between the peaks of the chromatogram is presented in formula (6). This formula describes the distance between the i-th and j-th peak.

$$d_{i,j} = (tr_i - tr_j)^2 \tag{6}$$

Based on the considerations mentioned above, a criterion for selecting the stationary layer for a given data set can be defined. The structure of the stationary phase - elements of the FS vector should be chosen so that for a given data vector, the sum of the distances between peaks is the largest; this relationship is expressed by the formula (7).

$$\begin{gathered} E(fs_1, fs_2, fs_3 \dots fs_M) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} d_{i,j} \\ \sum_{i=1}^{N} \sum_{j=i+1}^{N} (tr_i - tr_j)^2 \end{gathered} \tag{7}$$

In other words, the elements of the stationary phase should be selected so that the expression described in formula (7) representing the sum of the distances between peaks has the most significant value.

$$max(E(fs_1, fs_2, fs_3 \dots fs_M)) \tag{8}$$

The conditions presented in formulas (9) and (10) must be met to find the maximum of the function.

$$\begin{gathered} \frac{\partial E(fs_1, fs_2, fs_3 \dots fs_M)}{\partial fs_1} = 0 \\ \frac{\partial E(fs_1, fs_2, fs_3 \dots fs_M)}{\partial fs_2} = 0 \\ \dots \\ \frac{\partial E(fs_1, fs_2, fs_3 \dots fs_M)}{\partial fs_M} = 0 \end{gathered} \tag{9}$$

$$\begin{gathered} \frac{\partial^2 E(fs_1, fs_2, fs_3 \dots fs_M)}{\partial fs_1^2} < 0 \\ \frac{\partial^2 E(fs_1, fs_2, fs_3 \dots fs_M)}{\partial fs_2^2} < 0 \\ \dots \\ \frac{\partial^2 E(fs_1, fs_2, fs_3 \dots fs_M)}{\partial fs_M^2} < 0 \end{gathered} \tag{10}$$

To simplify further considerations and without losing the generality of the conclusions drawn, suppose the stationary phase consists of two elements M=2 and the number of substances for which we want to calculate the chromatogram is four N=4, then the expressions presented above will take the following form:

$$FS = (fs_1, fs_2) \tag{11}$$

$$S = \begin{bmatrix} s_{1,1}, s_{1,2}, \\ s_{2,1}, s_{2,2} \\ s_{3,1}, s_{3,2} \\ s_{4,1}, s_{4,2} \end{bmatrix} \tag{12}$$

$$\begin{aligned} E(fs_1, fs_2) &= \sum_{i=1}^{4} \sum_{j=i+1}^{4} d_{i,j} \\ &= d_{1,2} + d_{1,3} + d_{1,4} + d_{2,3} + d_{2,4} + d_{3,4} \\ &= \sum_{i=1}^{4} \sum_{j=i+1}^{4} \{F_{tr}(s_{i,1..M}, FS) - F_{tr}(s_{j,1..M}, FS)\}^2 \end{aligned} \tag{13}$$

The function we maximize for a given input set does not have a maximum. Below is a graph of this function for an example data set.

The presented graph shows that the function defined by formula (13) or (7) does not have a maximum in the function responsible for determining the retention time (4), but if the elements of the stationary phase vector have the same sign, the value of the function (13) is not limited. This means that the distances between individual peaks will increase proportionally as long as the values of the stationary phase elements increase and if all the stationary phase elements have the same sign. The existence of such a relationship is beneficial. Still, if we classify highly distorted vectors, the distances between the peaks of the classified chromatogram may differ significantly from the peaks of the chromatogram of the reference vector, which will result in incorrect classification. In this case, replace function (5) with a non-linear function. Another possible solution to the problem of classifying complex chromatograms is the implementation of mechanisms that occur in a different type of chromatography, i.e., affinity chromatography. [24], [25], [26], [27], [28], [29], [30].

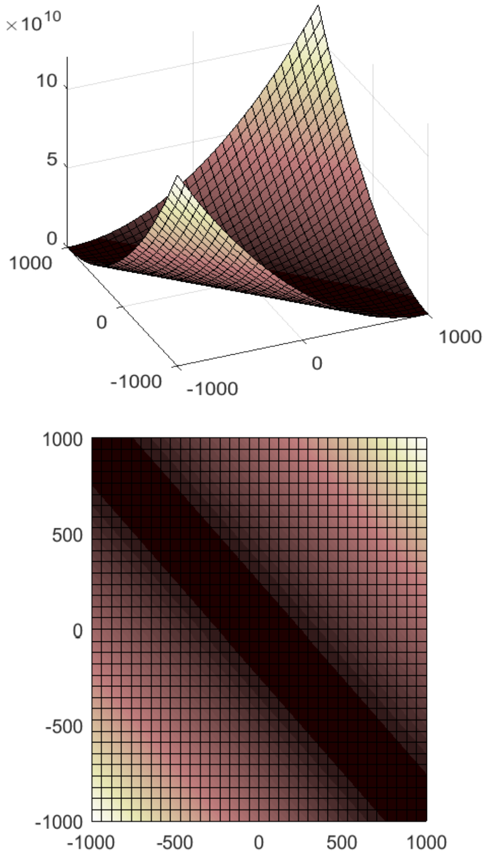


Fig. 8 Graph of the maximized function from formula (13) depending on the values of the elements of the stationary phase.

III. Results and Discussion

This chapter will present the algorithm's operation on an exemplary data set in which the dimension—the length of a single vector—has at least several hundred elements.

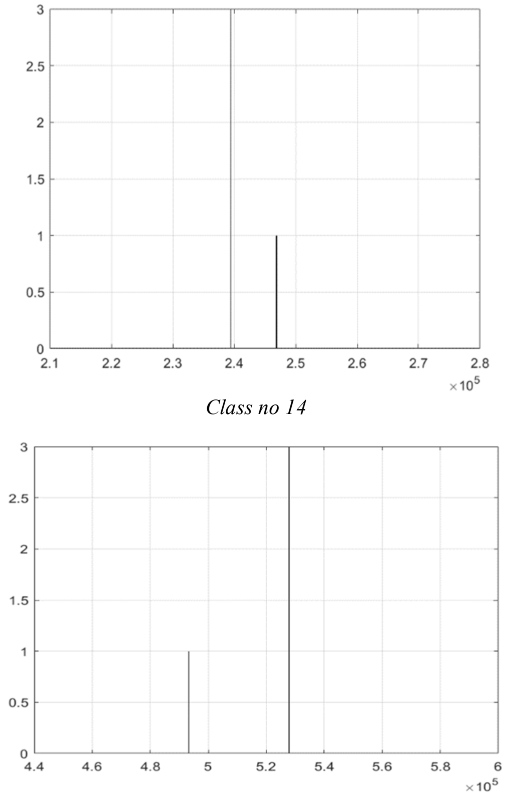
A. *Classification of a Homogeneous and One-class Dataset*

For this article, an "artificial" data set was created in pseudocode according to the formula presented below. As the presented algorithm shows, the training set consists of one hundred particles - vectors, each belonging to a separate class. The number of classes and vectors in the generated training dataset will be 100. However, the length of each vector-particle is 9425 elements.

```
aStep:=0.01
aEnd:=30
molecules:=[]
For i:=1:100
   A:=i
   f:=100/i
    molecules=[molecules { A*cos(f*(0:aStep:aEnd*pi) ) }]
end
```

Algorithm. 4 Generation of an artificial training dataset - program code in MATLAB

The result of processing the training set by the algorithm will be the creation of a set of chromatograms; each chromatogram will correspond to one vector from the training set. Figure 9 shows the chromatograms for two classes of the training set, which were created in the training process. On the horizontal axis of the presented graphs is the retention time. In contrast, on the vertical axis, it is the concentration of molecules that were created in the process of fragmentation of the input vector. In this case, the fragmentation level was 200, which means that the vectors that were processed by the chromatograph were divided into molecules containing 200 elements each - according to algorithm 1. For each created molecule, the retention time tr was calculated, and a chromatogram was created for each set of molecules as described in Algorithm 2.

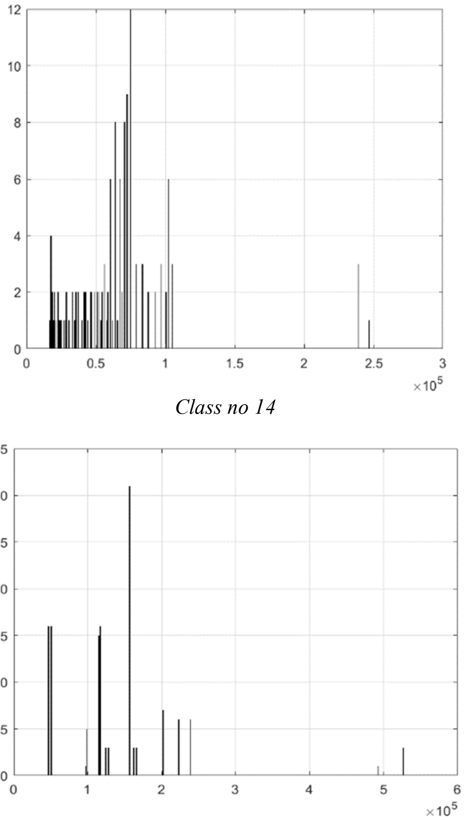


Class no 14

Class no 20

Fig. 9 Chromatogram for class no. 14 and 20 for fragmentation level 200.

In a situation where the fragmentation will be a multi-stage fragmentation, the chromatograms will contain a proportionally more significant number of peaks. Thus, the chromatograms will include a larger number of peaks. In a situation where the fragmentation will be a multi-stage fragmentation, the chromatograms will contain a proportionally more significant number of peaks. Thus, the chromatograms will include a larger number of peaks. Figure 6 shows the chromatogram of the same classes as before, but in this case, the division was multiple. Namely, the input vector was divided into 20-element, 50-element, and 200-element elements. You can see that new peaks have appeared in this case, but more importantly, the peaks that occurred with the single split have also been preserved.



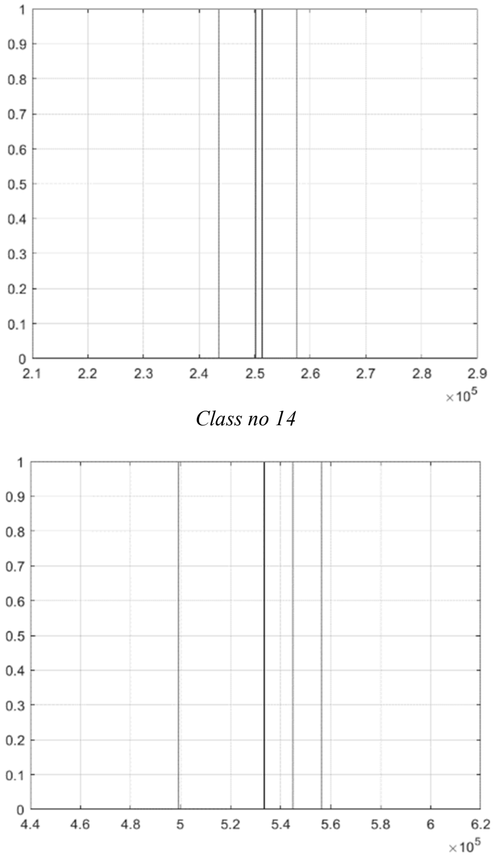
Class no 14

Class no 20

Fig. 10 Chromatogram for class no. 14 and 20 for fragmentation levels 20, 50, 200

The test set contained the same elements as the training set, but they were distorted; the level of distortion was 50%. Figure 11 shows the chromatograms of the test set for the same classes as before, but as mentioned, the data based on which these chromatograms were created were subjected to the distortion process. As for such a high level of distortion, the peaks in the presented chromatograms either have the same retention time that occurred in the chromatograms in Figure 9, or they differ slightly from these values.

The classification process, i.e., assigning an unknown chromatogram to a reference chromatogram, is carried out according to algorithm 3. In the discussed case, the classification process was successful. Table 1 presents the results of classification using the chromatographic data separation algorithm. As can be seen from the given data, the algorithm is relatively good at generalizing data because of a relatively high level of distortion of the testing data. In the case of data distortions of fifty percent, the percentage of correct classifications reached the level of 77.50% of correct answers.



Class no 14

Class no 20

Fig. 11 Chromatogram for class no. 14 and 20 for the level of fragmentation 200 with the level of distortion 50%

TABLE I
THE RESULTS OF THE CLASSIFICATION BY THE CHROMATOGRAPHIC ALGORITHM

No	The distortion level of the test data [%]	Fragmentation	Percentage of Correct Classifications [%]
1	50	50, 100	72.50
2	50	50,100, 150	67.50
3	50	50,100, 150, 200	72.50
4	50	20, 50, 100, 150, 200	77.50
5	25	50, 100	90.00
6	25	50, 100, 150	95.00

As the presented table also shows, the level of vector fragmentation has a significant impact on the level of correct answers. A higher level of vector fragmentation contributes to a few percent increase in the correct answers of the algorithm. The question arises of whether the presented algorithm is better at data classification tasks than other commonly known algorithms because no comparison of this algorithm with other data classification techniques has been included in this work. The reason for this is quite prosaic; the presented algorithm classifies vectors that have 9425 elements, which, in the case of artificial neural networks, it is practically impossible to classify vectors with such many elements without reducing the vector dimension, e.g., using mathematical statistics methods.

B. Classification of a Heterogeneous Dataset

In this section, an example of classifying a data set that is not homogeneous will be presented. This set's data vectors may have different lengths and be of different types, e.g., text, images, or numerical vectors. This algorithm makes this possible due to the assumption made regarding the structure of the data that is processed by the algorithm.

```
DataSet=[];
DataSet=[DataSet;{-10+sin(0:0.1:5*pi) }]
DataSet=[DataSet;{sin(0.5*(0:0.1:7*pi)) }]
DataSet=[DataSet;{5*cos(0.5*(0:0.1:6*pi)) }]
DataSet=[DataSet;{2000*sin(0:0.1:2*pi) }]
DataSet=[DataSet;{1:2000}]
DataSet=[DataSet;{20*sin(0.2*(0:0.1:5*pi)) }]
DataSet=[DataSet;{100*sign(sin(1.5*(0:0.1:9*pi))) }]
DataSet=[DataSet;{100*exp(0.01*(0:0.1:5*pi)).* sin((0:0.1:5*pi)) }
DataSet=[DataSet;{ 'Novel type of algorithm inspired by the mechanism of
chromatographic separation'}]
DataSet=[DataSet;{'Chromatography is a physicochemical separation
method in which the separated components are divided between two
phases: the stationary phase and the mobile phase move in a specific
direction. The different division of the mixture components between the two
phases causes the differentiation of the migration speed of the individual
components.' }]
```

Algorithm. 5 Generation of an artificial heterogeneous training dataset and testing dataset - program code in MATLAB

Algorithm 5 shows the definitions of the ten classes used to create the standard chromatograms. The individual input vectors are of different lengths. The test set was prepared similarly, except that, apart from the added noise, the lengths of individual vectors were modified.

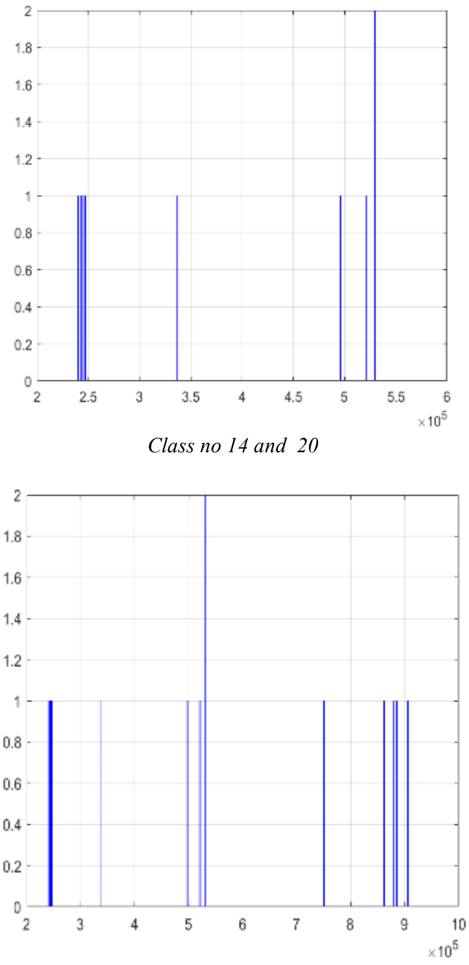
TABLE II
THE RESULTS OF THE CLASSIFICATION BY THE CHROMATOGRAPHIC ALGORITHM

No	The distortion level of the test data [%]	Fragmentation	Percentage of Correct Classifications [%]
1	0	10	83.37
2	50	10	66.67
3	0	20	91.37
4	50	20	74.32
5	0	30	99,89
6	50	40	74,65
7	0	40	100.0
8	50	40	75

The classification results are presented in Table 2. As can be seen, the algorithm did not always correctly classify in the absence of distortions in the testing set. This situation occurs when the level of data fragmentation is high, and it is related to the loss of information during the division of the vector into sub-vectors. In the case of reducing the level of fragmentation, the algorithm is classified flawlessly.

C. Classification of multi-class vectors

In this section, the results of classification with the presented algorithm of vectors whose contents belong to several classes at the same time will be presented. For example, suppose we have a vector containing data belonging to two classes, e.g., class 14 and class 20. Then, the chromatogram of such a vector will look as shown in Fig 12. As in the case of a real chromatographic system and in the case of the presented algorithm, it should be possible to identify a data vector that contains data belonging to many classes. Such a case will correspond to a real situation where the tested mixture consists of several chemical substances. This feature of the algorithm results from the adopted assumption that the vector that is processed by the presented algorithm in the first phase of the algorithm's operation is divided into smaller fragments, which corresponds to the process of creating a mixture, which is separated in the next steps of the algorithm.



Class no 14 and 20

Class no 14 20 and 26

Fig. 12 Chromatograms of vectors with a fragmentation level of 200 belonging to a) two classes, class 14 and class 20, and b) three classes: 14, 20,26

As can be seen from the presented figures, the chromatograms contain peaks appropriate for chromatograms belonging to class 14 and class 20. It follows that with the proposed data processing technique, it is possible to classify cases in which the classified vector belongs to several classes simultaneously. To present this algorithm property, as in the previous case, an artificial training set will be created, which will generate a set of standard chromatograms. The algorithm responsible for generating this data set is presented above (i.e., algorithm 4).

```
aStep:=0.05
aEnd:=30
molecules:=[]
For i:=1:100
```

```
5       A:=i
6       f:=100/i
7        molecules=[molecules, {  A*cos( f*(0:aStep:aEnd*pi) ) }]
8     end
9     mixture=[]
10    For  i:=1:50
11             mixture=[mixture, { molecules{i},  molecules{i+50} }]
12    end
```

Algorithm. 6 Generation of artificial multi-class testing set - program code in MATLAB (training dataset Algorithm 4)

The testing set, on the other hand, is created using algorithm 6. According to the presented algorithm, the vector that will be subjected to the classification process will simultaneously belong to two classes (algorithm 6—line 11). For example, vector 1 belongs to class one and fifty-first, etc. The task of the classifier (algorithm 3) will be to identify such a case by providing the number of classes to which fragments of the classified vector belong.

TABLE III
THE RESULTS OF THE CLASSIFICATION BY THE CHROMATOGRAPHIC ALGORITHM

No.	The distortion level of the test data [%]	Fragmentation	Correct Classifications [%]
1	0	5	100
2	25	10	83.33
3	25	20	76.67
4	0	35	81,0
5	50	35	70,09
6	0	50	50
7	50	50	40

In the case of this trial of the algorithm, the criterion for the correct classification of multi-class sets was that the correct classification is a classification in which all the classes to which the classified vector belongs have been correctly indicated.

As shown in the classification results presented in Table 3, in the case of multi-class data sets, i.e., two-class data sets, the presented algorithm performed correct classification at various levels. The level of classification correctness depended on two critical factors. Firstly, it relies on the level of fragmentation (algorithm 1) and the level of data distortion in the testing set. The higher the level of distortion and the lower the level of fragmentation, the number of correct classifications decreased. The reason for this unfavorable phenomenon is the construction of the stationary layer, equation (5); more considerable distortions result in more significant shifts relative to the reference chromatogram, and in such a case, the classification algorithm (algorithm 3) - line 4 may not consider such a distorted peak. Therefore, at a higher level of fragmentation, distortions do not play such a significant role as in the case of long fragments, when the substances in the mixture contain long chains of molecules. This phenomenon can be minimized by redefining the function that describes the interactions between the molecule and the stationary phase (equation 5). The table shows that if the data were not corrupted, the percentage of correct answers on the chromatographs was over ninety percent. Of course, in sets of three or four classes, the rate of correct answers will each be lower by several to a dozen or so percent.

IV. Conclusion

The article presents an algorithm for chromatographic data separation inspired by one of the analytical chemistry methods, resolution chromatography. Three algorithms used in chromatographic data separation have been proposed, constituting the chromatographic data separation process. Algorithm 1 is responsible for transforming a set of input vectors into a set of mixtures of substances. Algorithm 2 is the algorithm that is responsible for calculating the retention time. The third algorithm assigns the spectrum of the unknown substance, i.e., the input vector, to the chromatograms of the reference vectors.

The problems of selecting the stationary phase are presented in the following parts of the article. The classification process's quality depends on the vector's value, which is the stationary phase in the presented algorithm. It has been shown that in the case when the interaction between molecules and the stationary phase is described by formula (5), then if there is a correspondence of signs between all elements of the stationary phase, the distance between peaks will depend only on the values of the stationary phase elements. Increasing the value of the stationary phase elements will increase the distance between peaks, which will have a decisive impact on the correct operation of algorithm 3.

The paper presents the classification results of three different types of data sets. In the first case, a set classification process was carried out. The individual vectors contained approximately 10,000 elements. It is practically impossible to classify vectors of such dimensions directly using commonly known methods without reducing the dimension of the data. The second type of data set is a heterogeneous set containing various types of data, where, as in the first case, the input data vectors are suitably long. And for the rest, the third type of testing data set is a multi-class set. As shown in all the types mentioned above of datasets, the proposed classification mechanism performed relatively well. To improve the classification efficiency of the presented mechanism, it would be necessary to first algorithmize the problem of selecting the stationary phase considering nonlinear functions.

Based on the presented results, it can be assumed that the technique of chromatographic data separation can be successfully used in the processing of large data sets, where the data do not always have such features as a constant length of vectors, a relatively small number of elements in vectors, etc.

References

[1] D. Reinsel, J. Gantz, and J. Rydning, “Data Age 2025: The Evolution of Data to Life-Critical Don’t Focus on Big Data; Focus on the Data That’s Big Sponsored by Seagate The Evolution of Data to Life-Critical Don’t Focus on Big Data; Focus on the Data That’s Big,” 2017, Accessed: Feb. 06, 2024. [Online]. Available: www.idc.com

[2] M. Hilbert, “Big Data for Development: A Review of Promises and Challenges,” Development Policy Review, vol. 34, no. 1, pp. 135–174, Dec. 2015, doi: 10.1111/dpr.12142.

[3] A. Brabazon, M. O’neill, and S. Mcgarraghy, “Natural Computing Series Natural Computing Algorithms”, Accessed: Feb. 06, 2024. [Online]. Available: www.springer.com/series/

[4] S. D. Varhadi, V. A. Gaikwad, R. R. Sali, K. Chambalwar, and V. Kandekar, "A Short Review on: Definition, Principle and Applications of High Performance Liquid Chromatography Introduction," vol. 19, no. 2, pp. 628–634, 2020, Accessed: Feb. 06, 2024. [Online]. Available: www.ijppr.humanjournals.com

[5] Q.-H. Wan, *Mixed-Mode Chromatography Principles, Methods, and Applications*. Springer Singapore, Imprint: Springer, 2021.

[6] "Principles and Practice of Modern Chromatographic Methods," 2022, doi: 10.1016/c2019-0-03803-4.

[7] "Encyclopedia of Separation Science | ScienceDirect." Accessed: Feb. 06, 2024. [Online]. Available: https://www.sciencedirect.com/referencework/9780122267703/encyclopedia-of-separation-science

[8] A Research Agenda for Transforming Separation Science. National Academies Press, 2019. doi: 10.17226/25421.

[9] J. Martens, R. Bhushan, M. Sajewicz, and T. Kowalska, "Chromatographic Enantioseparations in Achiral Environments: Myth or Truth?," Journal of Chromatographic Science, vol. 55, no. 7, pp. 748–749, Apr. 2017, doi: 10.1093/chromsci/bmx031.

[10] M. Witting and S. Böcker, "Current status of retention time prediction in metabolite identification," Journal of Separation Science, vol. 43, no. 9–10, pp. 1746–1754, Apr. 2020, doi: 10.1002/jssc.202000060.

[11] G. J. Caluin, Dynamics of Chromatography Principles and Theory. CRC Press, 2017. doi: 10.1201/9781315275871.

[12] H. Schmidt-Traub, M. Schulte, and A. Seidel-Morgenstern, Eds., "Preparative Chromatography," Mar. 2020, doi: 10.1002/9783527816347.

[13] M. K. Gupta and P. K. Biswas, "Chromatography: Basic principle, types, and applications," Basic Biotechniques for Bioprocess and Bioentrepreneurship, pp. 173–182, 2023, doi: 10.1016/b978-0-12-816109-8.00010-6.

[14] "Chromatography: Definition, Working, and Importance in Various Industries." Accessed: Feb. 06, 2024. [Online]. Available: https://www.researchdive.com/blog/what-is-chromatography-how-does-it-work-and-where-is-it-used

[15] "Calculators| Chromatography Equations - MicroSolv Technology Corp MTC-USA." Accessed: Feb. 06, 2024. [Online]. Available: https://www.mtc-usa.com/calculators

[16] J. Pezzatti et al., "Implementation of liquid chromatography–high resolution mass spectrometry methods for untargeted metabolomic analyses of biological samples: A tutorial," Analytica Chimica Acta, vol. 1105, pp. 28–44, Apr. 2020, doi: 10.1016/j.aca.2019.12.062.

[17] "Introduction to Affinity Chromatography | Bio-Rad." Accessed: Feb. 06, 2024. [Online]. Available: https://www.bio-rad.com/en-pl/applications-technologies/introduction-affinity-chromatography?ID=LUSMJIDN

[18] J. R. Chapman, "Practical organic mass spectrometry : a guide for chemical and biochemical analysis," p. 330, 1993.

[19] A. Berthod, T. Maryutina, B. Spivakov, O. Shpigun, and I. A. Sutherland, "Countercurrent Chromatography in Analytical Chemistry," IUPAC Standards Online. De Gruyter, Mar. 09, 2016. doi: 10.1515/iupac.81.0005.

[20] T. O. Nicolescu, "Interpretation of Mass Spectra," Mass Spectrometry, Jun. 2017, doi: 10.5772/intechopen.68595.

[21] A. Knorr et al., "Computer-Assisted Structure Identification (CASI)—An Automated Platform for High-Throughput Identification of Small Molecules by Two-Dimensional Gas Chromatography Coupled to Mass Spectrometry," Analytical Chemistry, vol. 85, no. 23, pp. 11216–11224, Nov. 2013, doi: 10.1021/ac4011952.

[22] V. I. Babushok, "Chromatographic retention indices in identification of chemical compounds," TrAC Trends in Analytical Chemistry, vol. 69, pp. 98–104, Jun. 2015, doi: 10.1016/j.trac.2015.04.001.

[23] O. D. Sparkman, "Identification of essential oil components by gas chromatography/quadrupole mass spectroscopy Robert P. Adams," Journal of the American Society for Mass Spectrometry, vol. 16, no. 11, pp. 1902–1903, Nov. 2005, doi: 10.1016/j.jasms.2005.07.008.

[24] "Affinity Chromatography Principle, Procedure, Application, Advantages & Disadvantages - 2020 - YouTube." Accessed: Feb. 06, 2024. [Online]. Available: https://www.youtube.com/watch?v=zE0-F5TgpRs

[25] O. Jones, *Two-dimensional liquid chromatography principles and practical applications*. 2020. Accessed: Feb. 06, 2024. [Online]. Available: https://www.bookshop-santacruz.com/book/9789811561894

[26] "Learning by Simulations: Overlapping Peaks." Accessed: Feb. 06, 2024. [Online]. Available: https://www.vias.org/simulations/simusoft_peakoverlap.html

[27] L. Mondello, P. Q. Tranchida, P. Dugo, and G. Dugo, "Comprehensive two-dimensional gas chromatography-mass spectrometry: A review," Mass Spectrometry Reviews, vol. 27, no. 2, pp. 101–124, Jan. 2008, doi: 10.1002/mas.20158.

[28] A. Zaid, N. H. Hassan, P. J. Marriott, and Y. F. Wong, "Comprehensive Two-Dimensional Gas Chromatography as a Bioanalytical Platform for Drug Discovery and Analysis," Pharmaceutics, vol. 15, no. 4, p. 1121, Mar. 2023, doi: 10.3390/pharmaceutics15041121.

[29] M. Urh, D. Simpson, and K. Zhao, "Chapter 26 Affinity Chromatography," Guide to Protein Purification, 2nd Edition, pp. 417–438, 2009, doi: 10.1016/s0076-6879(09)63026-3.

[30] D. S. Hage, "Affinity Chromatography: A Review of Clinical Applications," 1999, Accessed: Feb. 06, 2024. [Online]. Available: https://academic.oup.com/clinchem/article/45/5/593/5643177.