

# RSC Advances



This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

*Accepted Manuscripts* are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. This *Accepted Manuscript* will be replaced by the edited, formatted and paginated article as soon as this is available.

You can find more information about *Accepted Manuscripts* in the [Information for Authors](#).

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard [Terms & Conditions](#) and the [Ethical guidelines](#) still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.



## A facile strategy applied to simultaneous qualitative-detection on multiple components of mixture samples: a joint study of infrared spectroscopy and multi-label algorithms on PBX explosives

Received 00th January 20xx,  
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/analyst

Minqi Wang<sup>a</sup>, Xuan He<sup>b</sup>, Qing Xiong<sup>a</sup>, Runyu Jing<sup>a</sup>, Yuxiang Zhang<sup>a</sup>, Zhining Wen<sup>a</sup>, Qifan Kuang<sup>a</sup>, Xuemei Pu<sup>a,\*</sup>, Menglong Li<sup>a,\*</sup> and Tao Xu<sup>b\*</sup>

We report a facile yet effective strategy of utilizing a combination of Fourier transform-infrared spectroscopy (FTIR) and multi-label algorithms, through which multi-components in polymer bonded explosives (PBXs) could be fast and simultaneously identified with high accuracy. The explosive components include 1,3,5,7-tetranitro-1,3,5,7-tetraazacyclo-octane (HMX), hexahydro-1,3,5-trinitro-1,3,5-triazine (RDX), 2,4,6-triamino-1,3,5-trinitrobenzene (TATB) and 2,4,6-trinitrotoluene (TNT) involved in single-component, binary-component and ternary-component PBXs. The train set contains 354 FTIR spectra of the explosives while the independent test set contains 84 ones. Two multi-label strategies (viz., data decomposition and algorithm adaptation) were adopted to construct the classification model with an objective of testing their efficiency in the multi-classification application. Principal component analysis (PCA) was applied to reduce the variables. Both the two algorithms exhibit excellent performance with 100 % accuracy for the training and the independent test sets. However, for real PBX samples, the performance of the algorithm adaptation strategy is sharply decreased to be 40% accuracy. But, it is noteworthy that the data decomposition strategy still achieves the accuracy of 100% for the real samples, exhibiting stronger robustness for the background interference and high promise in practice. The strategy proposed by the work would provide valuable information for advancing analytical methods in the explosive detection system and the other complicated samples.

### 1. Introduction

There is growing need to develop facile analysis methods to detect various explosives due to recent increase in terrorism activity.<sup>1-5</sup> In the past decades, many analytical methods based on instrumental techniques have been developed to determine the explosives,<sup>3-24</sup> involving in Raman spectroscopy,<sup>5-8</sup> Laser-induced breakdown spectroscopy,<sup>9-11</sup> Ion mobility spectrometry,<sup>12-14</sup> Mass spectrometry,<sup>15-17</sup> and Terahertz spectroscopy,<sup>18,19</sup> Gas chromatography,<sup>20-22</sup> High performance liquid chromatography<sup>23,24</sup> and so on. As reported, these instrumental methods are highly selective and sensitive. However, most of the devices are rather bulky, expensive and time-consuming, which impedes quick and on-line determination. Thus, it is highly desired to develop new methods or improve the existing techniques to enable faster, less expensive and simpler identification on the explosives.

As known, Fourier transform infrared spectroscopy (FTIR) is a relatively simple, rapid and nondestructive technique with

low running costs in qualitative determination and has been extensively used in many fields including explosives.<sup>5,25-27</sup> The technique provides fingerprint-like signatures of samples, resulting in diverse and complicated spectra. Consequently, it is difficult for FTIR to directly determine mixtures due to overlapped absorption band resulted from multiple components, which limits its further application in complicated systems, including mixture explosives.

Chemometrics methods<sup>28,29</sup> possess significant advantages in resolving the band overlaps from different components through mathematical separation instead of chemical separation. They have been successfully applied in the analysis of the complicated samples without pre-separation, including qualitative and quantitative determinations.<sup>30-33</sup> Recently, introduction of chemometrics methods to assist the determination of the explosives has aroused growing interest.<sup>34-39</sup> However, the previous works regarding the qualitative identification on the explosives mainly focused on the classification of single-component explosives with the aid of chemometrics methods,<sup>35,37-39</sup> while the mixture explosives were far less studied. With respect to the single-component explosives, the interference from multiple components of the mixture explosives would lead to more complicated signatures. Thus, it is necessary to introduce some other advanced chemometrics methods to deal with the simultaneous

<sup>a</sup> College of Chemistry, Sichuan University, Chengdu, People's Republic of China.  
Fax: +86-028-85412290; Tel.: +86-028-85412290; E-mail: xmpuscu@scu.edu.cn

<sup>b</sup> Institute of Chemical Materials, Chinese Academy of Engineering Physics, Mianyang, People's Republic of China.

\* Electronic Supplementary Information (ESI) available: See DOI: 10.1039/x0xx00000x

identification on the multiple components of the mixture explosives.

Pattern recognition methods used in the classification issue generally involve in two main strategies: single-label algorithm and multi-label one. The single-label classification deals with the instances that are associated with only one single label, while the multi-label classification<sup>40-42</sup> is an extension of traditional single label, in which the instances are associated with a number of labels simultaneously. Compared to the single-label classification, the multi-label learning exhibited more widely applications in real world, especially for multi-component identifications in many complicated cases, for example, text categorizations,<sup>43,44</sup> scene,<sup>45</sup> video annotation,<sup>46</sup> classifications in chemical systems, biological systems<sup>47-49</sup> and medical diagnosis.<sup>50,51</sup> In general, the multi-label identification can be performed via two algorithm strategies. One is data decomposition,<sup>52,53</sup> which splits the multi-label dataset into several single-label subsets and then combines the single-label derived from sub-classifiers on the subsets to give the identification result of the multicomponent samples. The other strategy is algorithm adaptation.<sup>54,55</sup> It uses the existing machine learning algorithm to tackle the multi-label prediction only by means of one single-classifier, which could simultaneously give multi-label information. The two strategies have been successfully applied in the multicomponent identification for some complicated systems.<sup>43-51</sup>

Although a few previous works regarding the qualitative analysis of the explosives partly involved in some mixture explosives,<sup>56,57</sup> they mainly concerned if these mixture samples were explosives or non-explosives, rather than the components comprising the explosives, thus, still belonging to the single-label classification. However, as known, the component compositions of the mixture explosives are closely associated with their explosion performances.

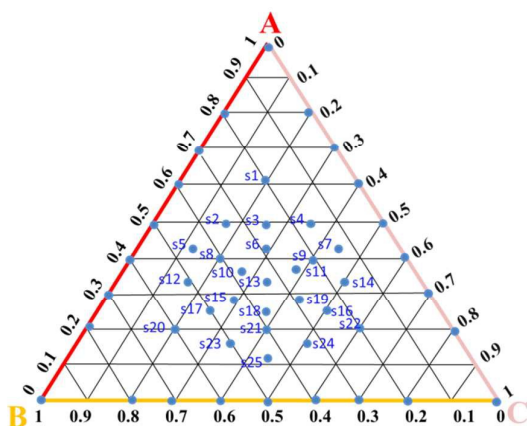
Based on the consideration above, we, herein, combined Fourier transform-infrared spectroscopy with the two multi-label strategies to develop a simple, quick and accurate method to realize simultaneous identification on the multiple components of the mixture explosives like polymer bonded explosives (PBXs). PBX is one kind of high-energy explosives that contains one to three energetic compounds (e.g., HMX, RDX, TATB, TNT) as main components and a small quantity of organic compounds (e.g., stabilizers, plasticizers, waxes, oils) as fillers.<sup>57,58</sup> It has been widely used due to their high energy density, mechanical strength and low sensitivity. Thereby, the identification on their components has emerged as an important task in industry and homeland security fields.<sup>6,59</sup> We selected and designed a series of single-component, binary-component and triple-component PBXs, involving in HMX (Cyclotetramethylene tetranitramine), RDX (cyclotrimethylenetrinitramine), TATB (triamino trinitrobenzene) and TNT (Trotyl), which are four of the most widely used secondary explosive ingredients in PBXs.<sup>60,61</sup> Consequently, 354 infrared spectra of the train dataset and 84 infrared spectra of the independent set were measured and constructed, which covered diverse composition proportion of

the energetic components. The two multi-label strategies mentioned above were used to establish the identification model between the component label of the explosives and the infrared spectral features of the PBXs by means of two well-accepted multi-label algorithms (viz., BR-SVM<sup>42,62</sup> and Rank-CVMz algorithm<sup>63</sup>). The results from the two algorithms were compared in order to assess their performances in the multi-label classification. Finally, the optimized models were applied to five real PBX samples. The results indicate that the multi-label algorithm based on the data decomposition strategy possess stronger robustness for the real samples than the algorithm adaptation strategy, at least for the mixture explosives. Thereby, a simple, quick and accurate method could be developed for the simultaneous detection on the multiple components of the PBXs by the FTIR spectroscopy coupled with the multi-label identification method, which potentially complement the explosive detection systems.

## 2. Materials and Methods

### 2.1 The Dataset Construction.

Considering that the real PBXs usually contain one to three energetic components, we constructed the data set by designing single-component, binary-component and ternary-component explosives based on pure HMX, RDX, TATB, and TNT analytes, which were provided by the Yinguang Chemical Plant, China. Similar to triangular mixture design strategy,<sup>64-66</sup> a series of mixture samples with different mass percentages were prepared, as illustrated by Fig.1. For example, the single-component explosives containing only one pure analyte were displayed at the poles of the triangle, in which A, B and C poles denote three of the four pure analytes. The binary-component samples containing seven different mass percentages of two pure analytes were displayed on the sides of the triangle. Six permutations of the two components from the four pure analytes lead to 42 formulations of the binary-component explosive samples. For the ternary-component explosive mixtures, we considered four permutations of three components from the four pure analytes, which were in general involved in real PBXs. Each of the four types of the ternary-component mixtures covers twenty-five formulations in the mass percentages, as displayed inside the triangle in Fig.1, leading to 100 ternary-component samples. Consequently, a total of 146 explosive samples were produced, which could efficiently represent the space of possible combinations of these ingredients, as reflected by Fig.1. The set of samples was randomly divided into the two data subsets. The train set was composed of 354 spectra with 3 spectra per sample for the 118 samples, which include 4 single-component explosives, 30 binary-component mixtures and 84 ternary-component mixtures. The independent test set was consisted of 84 spectra of 28 samples with 3 spectra per sample, including 12 binary-component mixtures and 16 ternary-component mixtures. In addition, five real polymer bonded explosives (PBXs) involved in the four analytes with inclusion of adhesives were also offered in order to validate application



**Fig.1** Illustration of sample distribution of the data set measured by FTIR. The A, B and C poles denote the three pure single-component explosive of the four analytes (HMX, RDX, TATB and TNT). The values 0 and 1 denote 0% and 100% mass percentage, respectively. The value between 0 and 1 represents the relative percentage in mass of the related analyte. For example, the sample labelled as 20 contains 20% component A, 60% component B, and 20% component C.

of the optimized models in practice. The data sets are available in the attachment file.

## 2.2 FT-IR Spectroscopic Measurements and Data Pre-processing.

A FT-IR spectrometer (Nicolet 5700) equipped with a diffuse reflection accessory was applied for the spectral scanning and the infrared spectra (IR) were collected in the Kubelka-Munk (K-M) mode. Samples were made by mixing explosives and potassium bromide (KBr) with 1:10, and then the samples were ground to fine powder and were put into the cuvette with a surface roughness. The wavenumbers were in the range of 4000-400  $\text{cm}^{-1}$  with the resolution 4.0  $\text{cm}^{-1}$ , and each spectrum was the average of 64 scans in order to obtain a good signal-to-noise ratio. Besides, the cuvette was emptied and treated with ethanol to avoid cross-contamination between different samples, and each sample was scanned 3 times parallelly to ensure the reproducibility.

The quality of infrared spectra strongly influences the performance of the analysis. Hence, pretreatment of infrared spectra should be taken carefully to get repeatable and reasonable results. In the work, baseline elimination and Savitzky-Golay smoothing with 5 segments were used to suppress baseline drift and additive noise. The pretreatments were performed using "OMNIC5.0" software implemented in the instrument. In addition, min-max normalization was applied to eliminate the random error and accelerate the calculation convergence as well as eliminate the effects of different samples thickness, which makes the data-processing more convenient.

## 2.3 Multivariate Analysis

**2.3.1 Principal Component Analysis (PCA).** Variable selection in multivariate analysis is a critical step to reduce collinearity and overlap. Principal component analysis (PCA)<sup>67-69</sup> is the

most widely used technique in variable extraction and dimension reduction. PCA transforms original variables into a few new variables called principal components (PCs). Each PC is a linear combination of the original variables, and they are orthogonal mutually. The differences and similarities among samples could be primarily visualized by means of projecting the data to a coordinate system defined by the two or three largest principal components. In general, the first principal component expresses the most variance in the data, and a few large principal components can explain most of the information.

**2.3.2 Support Vector Machine (SVM).** SVM<sup>69</sup> is a supervised classification algorithm which is adapted at the limited samples and less susceptible to over-fitting. SVM algorithm directly seeks the best balance between learning ability and model complexity. The input vectors are mapped from nonlinear space to higher dimensional space, a maximal separating hyperplane is constructed to separates two parallel hyperplanes on each class margin with a maximizes distance. The SVM software is implemented from the LIBSVM package<sup>70</sup> which can be freely downloaded from <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

**2.3.3 Multi-label Classification.** To date, there mainly exist two strategies for the multi-label classification: data decomposition and algorithm adaptation.<sup>42</sup> In general, the data decomposition strategy<sup>52,53,71</sup> splits the multi-label data set to one or more single-label subsets (binary or multi-class), and then trains a sub-classifier for each subset using an existing machine learning algorithm (for example, SVM, KNN or PLS-DA). Finally, all sub-classifiers are combined into an entire multi-label output. In the work, we adopted Binary Relevance (BR)<sup>62</sup> method to realize the multi-label classification of the data decomposition strategy since it is one well-accepted algorithm, which could transfer the multi-label learning problem to several binary classification ones in terms of one-vs-all strategy. Herein, the SVM algorithm was adopted in the BR strategy (labelled as BR-SVM). The algorithm adaptation strategy extends a specific multi-class algorithm to consider all component labels of every sample simultaneously.<sup>54,55,72,73</sup> Accordingly, some extended machine learning algorithms would be used in the algorithm adaptation strategy. In the work, we selected an extended CVM algorithm Rank-CVMz<sup>63</sup> (core vector machine with a zero label) as one representative, which adds a zero label as a benchmark label to distinguish related and unrelated marking. The Rank-CVMz was recently reported to have higher performance than six other multi-label algorithms<sup>63</sup> and could be achieved with the software in: [http://computer.njnu.edu.cn/Lab/LABIC/LABIC\\_Software.html](http://computer.njnu.edu.cn/Lab/LABIC/LABIC_Software.html).

**2.3.4 Six Instance-based Performance Measures.** Since the multi-label classification is more complicated than the single-label one, various evaluation measures have been proposed.<sup>63</sup> In this work, six popular and indicative measures were applied to evaluate the performance of the models: Hamming loss, accuracy, precision, recall, F1 and subset accuracy, values of which are in the range from 0 to 1. In general, an excellent model should achieve a smaller value for hamming loss, and larger values for the other five measures. For a contain dataset

$S = \{(X_1, Y_1), \dots, (X_p, Y_p), \dots, (X_m, Y_m)\}$  and the label sets  $Y = \{1, 2, \dots, Q\}$ , the six measures are estimated in terms of the formulas as follows:

$$\text{Hamming loss} = \frac{1}{m} \sum_{i=1}^m \frac{|Y_i \Delta Y_i^p|}{Q}$$

Hamming loss estimates the percentage of the labels that are incorrectly predicted for an instance, the  $Y_i^p$  means the predicted label set of the instance  $X_i$ , the  $Y_i$  means the real label set, and the  $\Delta$  stands for the symmetric difference between two label sets.

$$\text{Accuracy} = \frac{1}{m} \sum_{i=1}^m \frac{|Y_i \cap Y_i^p|}{|Y_i \cup Y_i^p|}$$

Accuracy evaluates the fraction of the correctly predicted labels out of the union of all predicted and true labels.

$$\text{Precision} = \frac{1}{m} \sum_{i=1}^m \frac{|Y_i \cap Y_i^p|}{|Y_i^p|}$$

Precision evaluates the percentage of the correctly predicted labels out of the predicted labels.

$$\text{Recall} = \frac{1}{m} \sum_{i=1}^m \frac{|Y_i \cap Y_i^p|}{|Y_i|}$$

Recall evaluates the percentage of the correctly predicted labels out of the true labels.

$$F1 = \frac{1}{m} \sum_{i=1}^m \frac{2|Y_i \cap Y_i^p|}{|Y_i| + |Y_i^p|}$$

F1 evaluates the harmonic mean between the precision and recall.

$$\text{Subset accuracy} = \frac{1}{m} \sum_{i=1}^m |\{1|Y_i = Y_i^p\}|$$

Subset accuracy evaluates the percentage of the label subsets that are predicted entirely correctly.

### 3. Results and Discussion

In this study, cyclotetramethylene tetranitramine (HMX), cyclotrimethylene trinitramine (RDX), triamino trinitrobenzene (TATB) and 2, 4, 6-trinitrotoluene (TNT) were provided as the energetic components of the explosives, Fig.2 shows their chemical structures and Fig.3 displays their infrared spectra. As shown, the four explosives have some similar functional groups like amine group, nitro group and benzene ring, which lead to the overlapped absorption regions ranging from 400  $\text{cm}^{-1}$  to 1700  $\text{cm}^{-1}$  and near 3000  $\text{cm}^{-1}$  in the spectra. Thus, it's necessary to introduce chemometrics method to assist experiments to simultaneously identify multiple components in the PBXs.

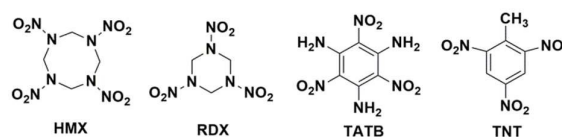


Fig.2 Molecular structures of HMX, RDX, TATB and TNT

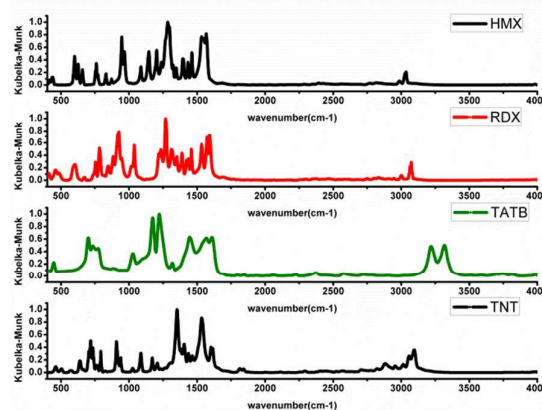


Fig.3 Infrared spectra of pure HMX, RDX, TATB and TNT explosives

#### 3.1 The construction of training models.

##### 3.1.1 Multi-label model based on data decomposition strategy (BR-SVM).

BR-SVM method identifies each component from the mixture explosives through building four sub-classifiers (HMX-classifier, RDX-classifier, TATB-classifier and TNT-classifier), as illustrated in Fig.4. Every sub-classifier was constructed on the basis of all data of the training set and performed one binary classification with one-vs-all strategy, in which the explosives containing the related component were served as the positive samples (labelled by 1) while the explosives without inclusion of the component were considered as negative samples (labelled by 0). For example, in the HMX sub-classifier, if the sample contained HMX component, either for the single-component explosive or for the multi-component explosive, it was assigned to the positive sample, or else it was classified into the negative sample. Finally, combined the labels derived from the four sub-classifiers, one final label set with inclusion of the four sub-classifier labels was output for each sample, revealing what components were contained in the explosive sample. For example, the label set of 1010 (see Fig.4) means that it contains HMX and TATB ingredients in the sample. Namely, it is one binary-component explosive formed by HMX and TATB.

**PCA analysis:** In order to avoid over-fitting and reduce the calculate complexity, it is of importance to reduce the number of features. PCA was applied to transform the original variables into a small number of new variables called principal components (PCs) with new coordinates called scores. The projection of the spectra onto the scores of the three principal components, which explain 89.04% information, could provide a visualization tool to check the inter-relationship between the different variables. Fig.5 shows the visual classification based on the first three principle components for the four binary classifiers, where the blue dots denote the positive samples

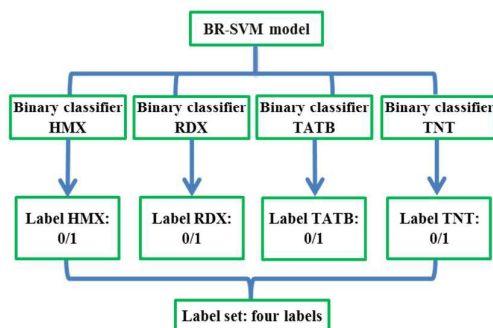


Fig.4 Schematic diagram of BR-SVM model based on the data decomposition strategy

while the red dots denote the negative samples. As can be seen from Fig.5, the TATB sub-classifier could differentiate the explosives containing TATB from the other ones without inclusion of TATB only based on the three principle components. However, the other three sub-classifiers cannot clearly discriminate the positive samples from the negative ones, implying that the information from the three principle components is not enough. In addition, we also evaluated the ability of the first three components of PCA to describe the FTIR spectra by means of a detail check their factor spectra derived from loadings of the three PCs (see Fig.6). As reflected by Fig.6, the two band regions (viz., 400-1600  $\text{cm}^{-1}$  and 2800-3400  $\text{cm}^{-1}$ ) could characterize the main differences between the three loading spectra. A detail comparison of the three loadings shows that with respect to the other two loadings, two significant peaks of 3217  $\text{cm}^{-1}$  and 3318  $\text{cm}^{-1}$  in the first loading, which are assigned to stretching vibrations of  $\text{NH}_2$  group, should mainly contribute to the clear discrimination between the explosives containing TATB and the other negative samples without TATB in Fig.5. In the second loading, it is noted that one 1285  $\text{cm}^{-1}$  peak assigned to N- $\text{NO}_2$  stretching vibration could characterize HMX and RDX components. The peak at 1356  $\text{cm}^{-1}$  assigned to  $\text{CH}_3$  bend should play a crucial role in discriminating between TNT and the other components. The observations indicate that the three principle components indeed contain important spectral information. However, as reflected by Fig.5, the simple classification based on the three principle components is far away from the clear discrimination between the different explosive components. Hence, PCA should be coupled with more advanced classification tool like SVM in order to perform the task.

**SVM Model optimization and performance:** In order to achieve a better predictive performance, each binary sub-classifier was trained by SVM based on all data of the training set and optimized separately. In this work, three important factors were considered in constructing the SVM-classifier: the number of PCs, regularization factors  $c$  and kernel scale factor  $g$ . As known, too little PCs are not capable to provide sufficient information for an accurate classification, leading to low predictive performance, while too many PCs may increase the complex of the model, resulting in over fitting. In addition,

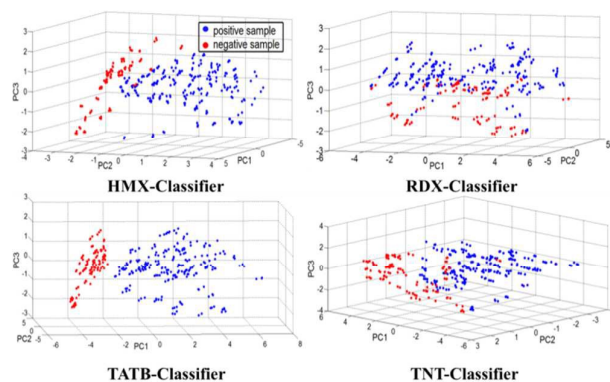


Fig.5 Visual classification based on the first three principle components for the four binary classifiers. Blue and red points represent positive and negative samples, respectively.

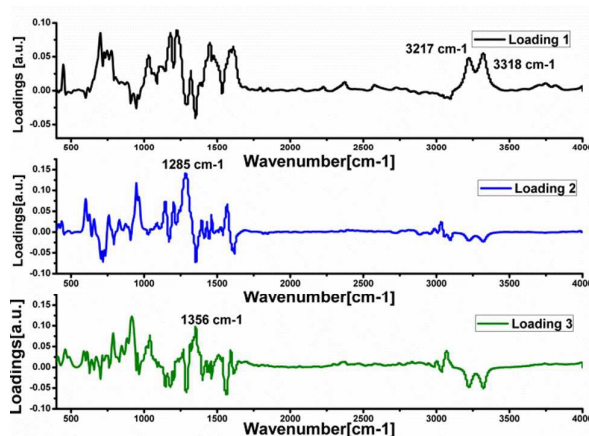


Fig.6 Factor spectra derived from the loadings of the first three principal components for training dataset

the smaller values of  $c$  and  $g$ , the better generalization and less possible to overfitting and overstudy. Five-fold cross validation and grid search technique were applied to obtain the optimal parameter set for the four sub-classifiers. Fig.7 shows the training performance at different  $c$  and  $g$  values varying with the number of PCs. Taking the impact of the three parameters together, we finally determine the optimized parameter set, as listed in Table 1. The optimized number of PCs are determined to be four, nine, three and seven for HMX sub-classifier, RDX one, TATB one and TNT one, which explain 95.69%, 99.37%, 89.04% and 98.79% of the total variance to the four classifiers, respectively. Thus, the picked PCs should be sufficient to represent most information of the spectra data. As expected, the optimized training model accurately identified every functional component of all samples in the training set and achieved 100% accuracy in every binary classifier. As a result, the Hamming loss of the model got the smallest value 0, and the other five measures (Accuracy, F1, Precision, Recall and Subset accuracy) got the maximum value 1.

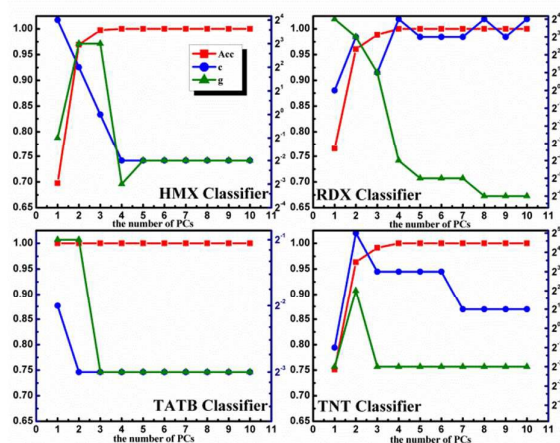


Fig.7 The tuning procedure of PCs,  $c$ ,  $g$  for the four binary-classifiers

Table 1 The training performance and parameter combination of each optimal sub-classifier for BR-SVM algorithm

Sub-Classifier	PCs <sup>a</sup>	Contribution <sup>b</sup>	Acc <sup>c</sup>
HMX	4	95.69%	100%
RDX	9	99.37%	100%
TATB	3	89.04%	100%
TNT	7	98.79%	100%

<sup>a</sup> the number of the picked PCs;

<sup>b</sup> the sum contribution of the picked PCs;

<sup>c</sup> the finally accuracy of each sub classifier.

**3.1.2 Multi-label model based on algorithm adaptation strategy (Rank-CVMz model).** Different from BR-SVM model, Rank-CVMz method is to use one extended classifier, rather than the four sub-classifiers, to identify all components in every sample through one label set with inclusion of the information of the four components, as illustrated in Fig.8. In the other words, the method averagely takes into account the variable features from the four components to optimize the single classifier. Also, the three important parameters ( $PCs$ ,  $c$  and  $g$ ) involved in SVM algorithm were optimized by means of a simple lazy-tuning procedure.<sup>63</sup> Similarly, the five-fold cross validation was used in the model construction. Fig.S1 in supplementary information displays the training performance at different parameters. Based on Fig. S1, we determined one set of optimized parameters with  $PCs$  of 7,  $g$  of 0.5 and  $c$  of 4. Using the optimized model, the high performance was achieved, as reflected by Subset accuracy of 99.72%, Hamming loss of 0.14%, Accuracy of 99.86%, Precision of 99.86%, Recall of 100% and F1 of 99.91%, exhibiting nearly 100% accuracy in identifying the components of the explosive mixtures for the training set.

### 3.2 Validation of the optimal Models by the independent test set

To validate the predictive performance of the two models, the optimized BR-SVM model and Rank-CVMz model were used to

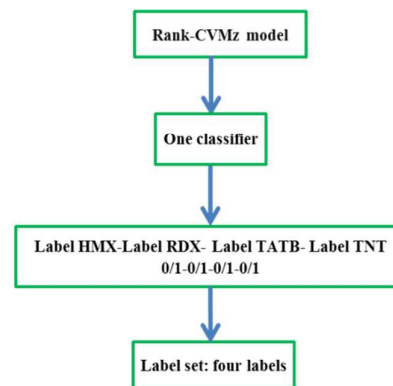


Fig.8 Schematic diagram of Rank-CVMz model based on the algorithm adaptation strategy

predict the components (HMX, RDX, TATB and TNT) of the mixture samples in the independent test set consisted of 84 spectra from 28 samples with 3 spectra per sample. Same as the performance of the training set, all the 84 spectra were accurately identified by the two models, exhibiting high prediction ability without over-fitting problem for simultaneous identification on HMX, RDX, TATB and TNT of the mixture samples in the independent set.

### 3.3 Application of the optimized models to the real samples

In order to test the applicability of the proposed methods in practice, the optimized BR-SVM model and Rank-CVMz models were further used to simultaneously identify HMX, RDX and TNT components in the real explosive samples. Five real explosives were provided, including two single-component HMX and RDX explosives, two binary-component explosives (HMX and RDX, HMX and TNT) and one ternary-component explosive (HMX, RDX and TNT). Their compositions were listed in Table 2. As shown in Table 2, the five real explosives contain a bit of additives besides the explosive components. The predictive results are also summarized in Table 2. Disappointedly, the Rank-CVMz correctly identified only two samples from the five real explosives. The three explosives PBXN-5, PBX 9407 and Octol were wrongly predicted to be the binary-component explosive composed of HMX and TNT, the ternary-component explosive containing HMX, RDX and TATB, and the mixture explosive formed by HMX, RDX and TNT, respectively. But, BR-SVM still retains 100% accuracy, exhibiting high potential in practice. One of the important reasons for the high prediction rate of BR-SVM should be attributed to the fact that it separately uses the sub-classifier to abstract the pertinent features of each component rather than considering the whole features of all components, thus, exhibiting strong ability of anti-disturbance from the additives in identifying the related component. Whereas for Rank-CVMz model, only one single-classifier was constructed through taking the entire features of the four energetic components into accounts, which may cause to some extent

Table 2 Prediction performance of the two optimized models for the real PBX samples

Sample	Composition (%)				BR-SVM		Rank-CVMz	
	HMX	RDX	TNT	Other	predicted label	explanation	predicted label	explanation
Termix	45	30	20	5 ammonium nitrate	HMX,RDX,TNT	Exact match	HMX,RDX,TNT	Exact match
PBX 9407	0	94	0	6 FPC461	RDX	Exact match	HMX,RDX,TATB	incomplete match
PBX	48	49	0	1.5 F2314 1.5 F2311	HMX,RDX	Exact match	HMX,RDX	Exact match
PBXN-5	95	0	0	5 Fluorelastomer	HMX	Exact match	HMX,TNT	incomplete match
Octol	67	0	30	3 F2314	HMX, TNT	Exact match	HMX,RDX,TNT	incomplete match

loss of individual characteristics for each component. Thus, with respect to BR-SVM model, it may be easier for the Rank-CVMz model to be influenced by the disturbance from the additives in the real explosives, leading to low identification accuracy. In addition, as reported, the algorithm adaptation strategy probably induces some complicated optimization problems since it simultaneously takes into account all labels.<sup>63,74</sup> Thus, the BR-SVM model exhibits higher application potential for the real explosive identification.

#### 4. Conclusions

This work combined the first time multi-label pattern recognition techniques with Fourier transform-infrared spectroscopy (FT-IR) to simultaneously detect the multiple components in the mixture explosives. The two main multi-label strategies (viz., data decomposition and algorithm adaptation) were used and compared in order to assess their performance in the multi-component recognition. The two strategies exhibited excellent performance with 100% accuracy for the training and independent data sets. However, the algorithm adaptation strategy based on Rank-CVMz model fails to accurately identify the five real PBX samples with only 40% accuracy, displaying weak anti-disturbance ability to the additives. But, the data decomposition strategy represented by the BR-SVM model still achieved 100% accuracy for the five real samples, exhibiting stronger robustness to eliminate disturbance from the background, thus, showing high potential for the explosive detection in practice. However, it should be noted that the data decomposition algorithm to some extent ignores correlation between multiple labels resulted from its one-against-all strategy that each label is treated individually. Thus, its performance may be weakened when there are strong correlations between multiple components for some mixture systems. In a whole, it is facile yet efficient for FT-IR spectrophotometry in combination with the multi-label algorithms to realize the simultaneous identification on the multiple components of PBXs in practice. Also, the strategy proposed by the work provides helpful information for advancing analysis method in other complicated systems.

#### Acknowledgements

This project is supported by the National Science Foundation of China (Grant No. U1230121 and 21273154).

#### References

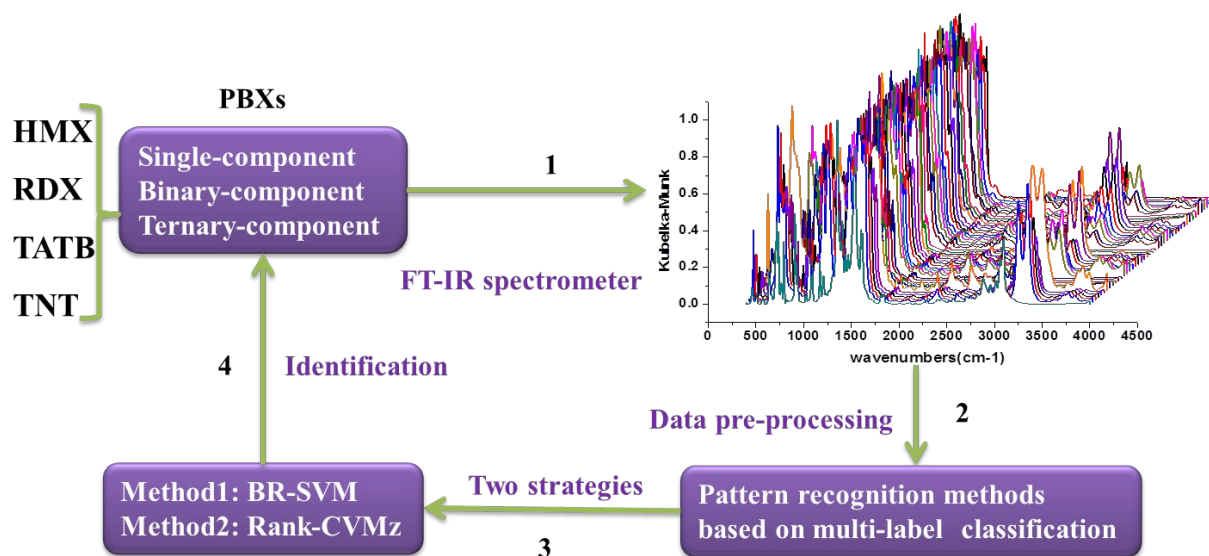
- J. H. Flexman, T. N. Rudakov, P. A. Hayes, N. Shanks, V. T. Mikhailitsevitch and W. P. Chisholm, Detection of bulk explosives: advanced techniques against terrorism in: H. Shubert, A. Kuznetsov, *NATO ASI Ser., Ser.II: Math. Phys. Chem.*, 2004, **138**, 113.
- M. A. Ivy, L. T. Gallagher, A. D. Ellington and E. V. Anslyn, Exploration of plasticizer and plastic explosive detection and differentiation with serum albumin cross-reactive arrays, *Chem. Sci.*, 2012, **3**, 1773-1779.
- J. S. Caygill, F. Davis and S. P. J. Higson, Current trends in explosive detection techniques, *Talanta*, 2012, **88**, 14-29.
- D. S. Moore, Instrumentation for trace detection of high explosives, *Rev. Sci. Instrum.*, 2004, **75**, 2499-2512.
- M. López-López and C. García-Ruiz, Infrared and Raman spectroscopy techniques applied to identification of explosives, *TrAC Trend. Anal. Chem.*, 2014, **54**, 36-44.
- C. Saint-Amans, P. Hébert, M. Doucet and T. de Resseguier, In-situ Raman spectroscopy and high-speed photography of a shocked triaminotrinitrobenzene based explosive, *J. Appl. Phys.*, 2015, **117**, 023102.
- C. Eliasson, N. A. Macleod and P. Matousek, Noninvasive detection of concealed liquid explosives using Raman spectroscopy, *Anal. Chem.*, 2007, **79**, 8185-8189.
- B. Zachhuber, G. Ramer, A. Hobro and B. Lendl, Stand-off Raman spectroscopy: a powerful technique for qualitative and quantitative analysis of inorganic and organic compounds including explosives, *Anal. Bioanal. Chem.*, 2011, **400**, 2439-2447.
- F. J. Fortes, J. Moros, P. Lucena, L. M. Cabalín and J. J. Laserna, Laser-induced breakdown spectroscopy, *Anal. Chem.*, 2012, **85**, 640-669.
- S. Sunku, M. K. Gundawar, A. K. Myakalwar, P. P. Kiran, S. P. Tewari and S. V. Rao, Femtosecond and nanosecond laser induced breakdown spectroscopic studies of NTO, HMX, and RDX, *Spectrochim. Acta. B*, 2013, **79**, 31-38.
- J. L. Gottfried, F. C. De Lucia Jr, C. A. Munson and A. W. Miziolek, 'Laser-Induced Breakdown Spectroscopy for Explosive Residue Detection: A Review of the Challenges, Recent Advances, and Future Prospects, *Anal. Bioanal. Chem.*, 2009, **395**, 283-300.
- R. G. Ewing, D. A. Atkinson, G. A. Eiceman and G. J. Ewing, A critical review of ion mobility spectrometry for the detection of explosives and explosive related compounds, *Talanta*, 2001, **54**, 515-529.
- G. R. Asbury, J. Klasmeier and H. H. Hill Jr, Analysis of explosives using electrospray ionization/ion mobility spectrometry ESI/IMS, *Talanta*, 2000, **50**, 1291-1298.



- 14 G. A. Buttigieg, A. K. Knight, S. Denson, C. Pommier and M. B. Denton, Characterization of the explosive triacetone triperoxide and detection by ion mobility spectrometry, *Forensic Sci. Int.*, 2003, **135**, 53-59.
- 15 A. N. Martin, G. R. Farquar, E. E. Gard, M. Frank and D. P. Ferguson, Identification of high explosives using single-particle aerosol mass spectrometry, *Anal. Chem.*, 2007, **79**, 1918-1925.
- 16 R. GrahamáCooks, Direct, trace level detection of explosives on ambient surfaces by desorption electrospray ionization mass spectrometry, *Chem. Commun.*, 2005, **15**, 1950-1952.
- 17 J. J. Brady, E. J. Judge and R. J. Levis, Identification of explosives and explosive formulations using laser electrospray mass spectrometry, *Rapid. Commun. Mass. Sp.*, 2010, **24**, 1659-1664.
- 18 J. Chen, Y. Chen, H. Zhao, G. J. Bastiaans and X. C. Zhang, Absorption coefficients of selected explosives and related compounds in the range of 0.1-2.8 THz, *Opt. Express*, 2007, **15**, 12060-12067.
- 19 M. R. Leahy-Hoppa, M. J. Fitch and R. Osiander, Terahertz spectroscopy techniques for explosives detection, *Anal. Bioanal. Chem.*, 2009, **395**, 247-257.
- 20 M. Nambayah and T. I. Quickenden, A quantitative assessment of chemical techniques for detecting traces of explosives at counter-terrorist portals, *Talanta*, 2004, **63**, 461-467.
- 21 S. A. Barshick and W. H. Griest, Trace analysis of explosives in seawater using solid-phase microextraction and gas chromatography/ion trap mass spectrometry, *Anal. Chem.*, 1998, **70**, 3015-3020.
- 22 C. J. Miller, G. Elias, N. C. Schmitt and C. Rae, Identification of Explosives from Porous Materials: Applications Using Reverse Phase High Performance Liquid Chromatography and Gas Chromatography, *Sens. Imag. Int. J.*, 2010, **11**, 61-75.
- 23 L. Čapka, Z. Večeřa, P. Mikuška, J. Šesták, V. Kahle and A. Bumbová, A portable device for fast analysis of explosives in the environment, *J. Chromatogr. A*, 2015, **1388**, 167-173.
- 24 D. Gaurav, A. K. Malik and P. K. Rai, High-performance liquid chromatographic methods for the analysis of explosives, *Crit. Rev. Anal. Chem.*, 2007, **37**, 227-268.
- 25 A. Banas, K. Banas, M. Bahou, H. O. Moser, L. Wen, P. Yang, Z. J. Li, M. Cholewa, S. K. Lim and C. H. Lim, Post-blast detection of traces of explosives by means of Fourier transform infrared spectroscopy, *Vib. Spectrosc.*, 2009, **51**, 168-176.
- 26 J. Akhavan, Analysis of high-explosive samples by Fourier transform Raman spectroscopy, *Spectrochim. Acta. A*, 1991, **47**, 1247-1250.
- 27 Y. Mou and J. W. Rabalais, Detection and Identification of Explosive Particles in Fingerprints Using Attenuated Total Reflection - Fourier Transform Infrared Spectromicroscopy, *J. Forest. Sci.*, 2009, **54**, 846-850.
- 28 N. Kumar, A. Bansal, G. S. Sarma and R. K. Rawal, Chemometrics tools used in analytical chemistry: An overview, *Talanta*, 2014, **123**, 186-199.
- 29 E. Szymańska, J. Gerretzen, J. Engel, B. Geurts, L. Blanchet and L. M. Buydens, Chemometrics and qualitative analysis have a vibrant relationship, *TrAC Trend. Anal. Chem.*, 2015, **69**, 34-51.
- 30 B. C. Deng, Y. H. Yun, Y. Z. Liang and L. Z. Yi, A novel variable selection approach that iteratively optimizes variable space using weighted binary matrix sampling, *Analyst*, 2014, **139**, 4836-4845.
- 31 Y. Liu, Y. Ning, W. Cai and X. Shao, Micro-analysis by near-infrared diffuse reflectance spectroscopy with chemometric methods, *Analyst*, 2013, **138**, 6617-6622.
- 32 Y. J. Yu, H. L. Wu, J. F. Niu, J. Zhao, Y. N. Li, C. Kang and R. Q. Yu, A novel chromatographic peak alignment method coupled with trilinear decomposition for three dimensional chromatographic data analysis to obtain the second-order advantage, *Analyst*, 2013, **138**, 627-634.
- 33 N. Qi, Z. Zhang, Y. Xiang and P. D. B. Harrington, Locally linear embedding method for dimensionality reduction of tissue sections of endometrial carcinoma by near infrared spectroscopy, *Anal. Chim. Acta*, 2012, **724**, 12-19.
- 34 F. C. De Lucia, J. L. Gottfried, C. A. Munson and A. W. Miziolek, Multivariate analysis of standoff laser-induced breakdown spectroscopy spectra for classification of explosive-containing residues, *Appl. Optics*, 2008, **47**, G112-G121.
- 35 J. Serrano, J. Moros, C. Sanchez, J. Macías and J. J. Laserna, Advanced recognition of explosives in traces on polymer surfaces using LIBS and supervised learning classifiers, *Anal. Chim. Acta.*, 2014, **806**, 107-116.
- 36 J. Moros, J. Serrano, C. Sanchez, J. Macias and J. J. Laserna, New chemometrics in laser-induced breakdown spectroscopy for recognizing explosive residues, *J. Anal. Atom. Spectrom.*, 2012, **27**, 2111-2122.
- 37 J. Hwang, N. Choi, A. Park, J. Q. Parkb, J. H. Chung, S. Baek, S. G. Cho, S. J. Baek and J. Choo, Fast and sensitive recognition of various explosive compounds using Raman spectroscopy and principal component analysis, *J. Mol. Struct.*, 2013, **1039**, 130-136.
- 38 K. Banas, A. Banas, H. O. Moser, M. Bahou, W. Li, P. Yang, M. Cholewa and S. K. Lim, Multivariate analysis techniques in the forensics investigation of the postblast residues by means of fourier transform-infrared spectroscopy, *Anal. Chem.*, 2010, **82**, 3038-3044.
- 39 R. Jimenez-Perez, L. Elie, M. Baron, J. Gonzalez-Rodriguez and Gonzalez-Rodriguez, Design of a virtual sensor data array for the analysis of RDX, HMX and DMNB using metal-doped screen printed electrodes and chemometric analysis, *Int. J. Electrochem. Sc.*, 2013, **8**, 3279-3289.
- 40 G. Tsoumakas and I. Katakis, Multi-label classification: An overview, Dept. of Informatics, Aristotle University of Thessaloniki, Greece, 2006, pp.64-74.
- 41 G. Tsoumakas, E. Spyromitros-Xioufis, J. Vilcek and I. Vlahavas, Mulan: A java library for multi-label learning, *J. Mach. Learn. Res.*, 2011, **12**, 2411-2414.
- 42 M. L. Zhang and Z. H. Zhou, A review on multi-label learning algorithms, *IEEE. T. Knowl. Data. En.*, 2014, **26**, 1819-1837.
- 43 F. Brucker, F. Benites and E. Sapozhnikova, Multi-label classification and extracting predicted class hierarchies, *Pattern Recogn.*, 2011, **44**, 724-738.
- 44 R. E. Schapire and Y. Singer, BoostText: A boosting-based system for text categorization, *Mach. Learn.*, 2000, **39**, 135-168.
- 45 M. R. Boutell, J. Luo, X. Shen and C. M. Brown, Learning multi-label scene classification, *Pattern Recogn.*, 2004, **37**, 1757-1771.
- 46 J. Wang, Y. Zhao, X. Wu and X. S. Hua, A transductive multi-label learning approach for video concept detection, *Pattern Recogn.*, 2011, **44**, 2274-2286.
- 47 Y. L. Wang, R. Y. Jing, Y. P. Hua, Y. Y. Fu, X. Dai, L. Q. Huang and M. L. Li, Classification of multi-family enzymes by multi-label machine learning and sequence-based descriptors, *Anal. Methods.*, 2014, **6**, 6832-6840.
- 48 E. A. Tanaka, S. R. Nozawa, A. A. Macedo and J. A. Baranauskas, A multi-label approach using binary relevance and decision trees applied to functional genomics, *J. Biomed. Inform.*, 2015, **54**, 85-95.
- 49 H. L. Zou and X. Xiao, A new multi-label classifier in identifying the functional types of human membrane proteins, *J. Membrane Biol.*, 2015, **248**, 179-186.
- 50 H. Z. Wang, X. Liu, B. Lv, F. Yang and Y. Z. Hong, Reliable multi-label learning via conformal predictor and random

- forest for syndrome differentiation of chronic fatigue in traditional chinese medicine, *PLoS One*, 2014, **9**, e99565.
- 51 G. P. Liu, G. Z. Li, Y. L. Wang and Y. Q. Wang, Modelling of inquiry diagnosis for coronary heart disease in traditional Chinese medicine by using multi-label learning, *BMC Complem. Altern. M.*, 2010, **10**, 37-48.
- 52 E. Montañes, R. Senge, J. Barranquero, J. R. Quevedo, J. J. del Coz and E. Hullermeier, Dependent binary relevance models for multi-label classification, *Pattern Recogn.*, 2014, **47**, 1494-1508.
- 53 J. Read, B. Pfahringer, G. Holmes and E. Frank, Classifier chains for multi-label classification, *Mach. Learn.*, 2011, **85**, 333-359.
- 54 S. Abe, Fuzzy support vector machines for multilabel classification, *Pattern Recogn.*, 2015, **48**, 2110-2117.
- 55 M. L. Zhang and Z. H. Zhou, ML-KNN: A lazy learning approach to multi-label learning, *Pattern Recogn.*, 2007, **40**, 2038-2048.
- 56 H. J. Im, B. C. Song, Y. J. Park and K. Song, Classification of materials for explosives from prompt gamma spectra by using principal component analysis, *Appl. Radiat. Isotopes.*, 2009, **67**, 1458-1462.
- 57 X. Cetó, A. M. O'Mahony, J. Wang and M. del Valle, Simultaneous identification and quantification of nitro-containing explosives by advanced chemometric data treatment of cyclic voltammetry at screen-printed electrodes, *Talanta*, 2013, **107**, 270-276.
- 58 Z. B. Zhou, P. W. Chen, F. L. Huang and S. Q. Liu, Experimental study on the micromechanical behaviour of a PBX simulant using SEM and digital image correlation method, *Opt. Laser. Eng.*, 2011, **49**, 366-370.
- 59 S. Maurer, R. Makarow, J. Warmer and P. Kaul, Fast testing for explosive properties of mg-scale samples by thermal activation and classification by physical and chemical properties, *Sensor. Actual. B-chem.*, 2015, **215**, 70-76.
- 60 C. M. Lin, J. H. Liu, F. Y. Gong, G. Y. Zeng, Z. Huang, L. P. Pan, J. H. Zhang and S. J. Liu, High-temperature creep properties of TATB-based polymer bonded explosives filled with multi-walled carbon nanotubes, *RSC Adv.*, 2015, **5**, 21376-21383.
- 61 S. Babaei and A. Beiraghi, Micellar extraction and high performance liquid chromatography-ultra violet determination of some explosives in water samples, *Anal. Chim. Acta.*, 2010, **662**, 9-13.
- 62 G. Tsoumakas, I. Katakis and I. Vlahavas, Mining multi-label data//Data mining and knowledge discovery handbook. Springer US, 2010, pp. 667-685.
- 63 J. Xu, Multi-label core vector machine with a zero label, *Pattern Recogn.*, 2014, **47**, 2542-2557.
- 64 T. Naes, T. Isaksson and B. Kowalski, Locally weighted regression and scatter correction for near-infrared reflectance data, *Anal. Chem.*, 1990, **62**, 664-673.
- 65 W. Saeys, K. Beullens, J. Lammertyn, H. Ramon and T. Naes, Increasing robustness against changes in the interferent structure by incorporating prior information in the augmented classical least-squares framework, *Anal. Chem.*, 2008, **80**, 4951-4959.
- 66 J. El Haddad, F. de Miollis, J. B. Sleiman, L. Canioni, P. Mounaix and B. Bousquet, Chemometrics applied to quantitative analysis of ternary mixtures by Terahertz spectroscopy, *Anal. Chem.*, 2014, **86**, 4927-4933.
- 67 K. L. Diehl and E. V. Anslyn, Array sensing using optical methods for detection of chemical and biological hazards, *Chem. Soc. Rev.*, 2013, **42**, 8596-8611.
- 68 P. C. Jurs, G. A. Bakken and H. E. McClelland, Computational methods for the analysis of chemical sensor array data from volatile analytes, *Chem. Rev.*, 2000, **100**, 2649-2678.
- 69 X. W. Feng, Q. H. Zhang, P. S. Cong, Z. L. Zhu, Preliminary study on classification of rice and detection of paraffin in the adulterated samples by Raman spectroscopy combined with multivariate analysis, *Talanta*, 2013, **115**, 548-555.
- 70 C. C. Chang and C. J. Lin, LIBSVM: A library for support vector machines, *ACM T. Intel. Syst. Tec. TIST*, 2011, **2**, 27.
- 71 K. Trohidis, Multi label classification of stack binary relevance models for multi-label classifiers, *Eurasip J. Audio. SPEE.*, 2011, **4**, 4-6.
- 72 J. Xu, An efficient multi-label support vector machine with a zero label, *Expert Syst. Appl.*, 2012, **39**, 2894-4796.
- 73 G. Tsoumakas, I. Katakis and I. Vlahavas, Random k-labelsets for multilabel classification, *IEEE. T. Knowl. Data. En.*, 2011, **23**, 1079-1089.
- 74 S. M. Liu and J. H. Chen, A multi-label classification based approach for sentiment classification, *Expert. Syst. Appl.*, 2015, **42**, 1083-1093.

## Graphical Abstract



We combined infrared spectroscopy with multi-label algorithms to propose a facile yet efficient strategy to realize simultaneous qualitative-detection on multiple components of mixture explosives without pre-separation.