Novel Similarity Methods Evaluation and Feasible Application for Pharmaceutical Raw Material Identification with Near-Infrared Spectroscopy

Hui Zhang, Lian Li, Shuang Quan, Weilu Tian, Kefan Zhang, Lei Nie,* and Hengchang Zang*

ABSTRACT: Raw material identification (RMID) is necessary and important to fulfill the quality and safety requirements in the pharmaceutical industry. Near-infrared (NIR) spectroscopy is a rapid, nondestructive, and commonly used analytical technique that could offer great advantages for RMID. In this study, two brand new similarity methods S1 and S2, which could reflect the similarity from the perspective of the inner product of the two vectors and the closeness with the cosine of the vectorial angle or correlation coefficient, were proposed. The ability of $u$ and $v$ factors to distinguish the difference between small peaks was investigated with the spectra of NIR. The results showed that the distinguishing ability of $u$ is greater than $v$, and the distinguishing ability of S2 is greater than S1. Adjusting exponents $u$ and $v$ in these methods, which are variable and configurable parameters greater than 0 and less than infinity, could identify small peaks in different situations. Meanwhile, S1 and S2 could rapidly identify raw materials, suggesting that the on-site and in situ pharmaceutical RMID for large-volume applications can be highly achievable. The methods provided in this study are accurate and easier to use than traditional chemometric methods, which are important for the pharmaceutical RMID or other analysis.

INTRODUCTION

The essential role of active pharmaceutical ingredient identification and raw material identification (RMID) in controlling pharmaceutical manufacturing products and the final product has been firmly approved by the U.S. Food and Drug Administration (FDA), which claims that “each lot of components, drug product containers, and closures shall be withheld from use until the lot has been sampled, tested, or examined, as appropriate, and released for use by the quality control unit” in FDA Title 21 Code of Federal Regulations (CFR) 211.84. In addition, guidelines confirmed that the need for the identification tests was also proposed by the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). Strictly speaking, many regulatory bodies require the inspection of every barrel in every shipment of materials used in pharmaceutical drugs (100% inspection) to avoid the contaminated or mislabeled materials risks.

In general, pharmaceutical RMID has relied on laboratory-based analytical techniques such as chromatography, titrations, and mid-infrared spectroscopy, which are time-consuming and laborious. It is worth mentioning that the suitability of near-infrared (NIR) spectroscopy, a rapid and nondestructive tool based on combinations and overtone of hydrogen bonds (i.e., C–H, N–H, and O–H), which could be performed through the packaging noninvasively and rapidly, is particularly getting more popular. The NIR for application in pharmaceutical testing has been addressed in the U.S. Pharmacopoeia (Chapter 1119) and the European Pharmacopoeia (Chapter 2.2.40). In addition, NIR spectroscopy has also gained wide acceptance in the pharmaceutical industry including the drying unit, granulation unit, blending unit, and so on in recent years.

Pharmaceutical RMID with NIR spectroscopy relies on the use of pattern recognition methods (PRMs), which can be divided into three major categories, including unsupervised, supervised, and spectra searching methods. Unsupervised methods could search for clustering in an N-dimensional space without knowing the class to which the sample belongs, such as cluster analysis and hierarchical cluster analysis (HCA). Supervised methods could assign samples to specific, previously known classes after prior training the system and obtaining optimal boundaries. Commonly used supervised methods are discriminant analysis (DA), the k-nearest neighbor (KNN), and the support vector machine (SVM).

Received: August 17, 2020
Accepted: October 12, 2020
Published: November 10, 2020
Spectra searching methods use the standardized spectrum as a reference to compare with the sample’s spectrum by similarity. The similarity is the main evaluation parameter that was taken as the extent to which a spectrum is identical to another. The correlation coefficient and cosine mainly reflect the sample similarity based on the similarity degree to the reference spectra, while the distance reflects the sample similarity based on the different degrees to the reference spectra. Here, we describe two new methods, S1 and S2, that combine spectral distance with the cosine method and introduce the sensitivity factors and as a powerful alternative for the spectra searching methods.

Through the use of several data sets of microcrystalline cellulose (MCC) PH101 and PH102, the performance of S1 and S2 was evaluated. MCC, a linear polysaccharide substance bound by β-1,4-glucoside bonds with a polymerization degree at about 3000–10,000, is not only used as a binder or diluent in oral tablets and capsules but also widely used for wet granulation and direct compression in the pharmaceutical process. Avicel PH 101 and Avicel PH 102 are the two most common grades of MCC used in tableting, which have very similar NIR spectra due to their small differences in properties. Moreover, five kinds of raw materials often utilized in pharmaceutical preparations were adopted to verify the feasibility of S1/S2 in RMID. The similarity results were calculated by adjusting the sensitivity factors and further compared with the results calculated by the classical correlation coefficient, cosine θ, partial least squares-discriminant analysis (PLS-DA), support vector machines (SVM), and soft independent modeling of class analogy (SIMCA).

The main purpose of this research was to provide alternative and convenient methods for RMID that could identify the raw material with high sensitivity and specificity. What needs to be clarified is, although the research in this paper focuses on pharmaceutical materials, the similarity methods described can be applied to other systems where spectral correlation methods are used.

RESULTS AND DISCUSSION

Comparison of Distinguishing Ability of $u/v$. The value of sensitive factor $u/v$ is of high importance for increasing or decreasing the distinguishing ability of S1 and S2. Theoretically, increasing the value of $u$ or $v$ can amplify the difference between the two spectra and reduce the similarity. Further explanation, lower similarity corresponds to better distinguishing ability. To investigate and compare the distinguishing ability of each factor in both methods, the similarity was calculated by mean spectra of PH101 and PH102 at three conditions, including (i) varying $u$ when $v = 1$, (ii) varying $v$ with $u = 1$, and (iii) varying $u$ and $v$ simultaneously from 1 to 100 with an interval of 1. Similarity results are shown in Figure 1a,b.

Figure 1a shows the similarity results acquired with S1. It could be seen that the increase of $u/v$ individually or simultaneously caused an almost linearly decrease in the similarity. The similarity at the same abscissa displayed that the distinguishing ability of factor $u$ is better than $v$, factors changing simultaneously better than those of $u/v$ changing individually. Figure 1b displays similarity results acquired with S2, showing a similar linear downward trend with S1, while the ordinate value is lower than S1 under the same conditions, indicating that S2 owes a better distinguishing ability.

To further analyze the distinguishing ability for PH101 and PH102 of each factor, the rate of change of similarity including $\Delta S1/\Delta u$, $\Delta S1/\Delta v$, $\Delta S2/\Delta u$, and $\Delta S2/\Delta v$ was calculated. Then, the data in Figure 1c were plotted to visualize the results and it indicated that $\Delta S2/\Delta u > \Delta S2/\Delta v > \Delta S1/\Delta u > \Delta S1/\Delta v$. Therefore, S2 has a better distinguishing ability than S1 and $u$ is better than $v$, which is consistent with the results in Figure 1a,b.

Investigation of Distinguishing Ability to Small Peaks. Absorption peak difference, including the peak height and width, of two NIR spectra, is the fundamental reason to be distinguished by any discriminative methods. To figure out the distinguishing ability of S1 and S2 for two spectra with a difference, the similarity between the spectrum of PH101 with and without small peaks was investigated. All peaks were simulated with a Gaussian function based on the mean spectrum of PH101 at 7000 cm$^{-1}$, where obvious bands of first overtones of the O–H stretching modes of cellulose are shown. Three aspects were studied, with 10 small peaks of (i) the peak height varying from 0.1 to 1 at a constant peak width, which was defined as the wide peak group, (ii) the same peak height but a narrower peak width to the wide peak group, and (iii) the constant peak area generated by varying sigma in the Gaussian function, defined as the constant peak area group. The simulated spectra and local enlarged images are all shown in Figure 2. It could be seen from Figure 2b,c that the maximum absorbance in wide and narrow groups increases regularly with the constant peak width. Nevertheless, the highest absorbance increases with the narrowing peak width in the constant area group, as depicted in Figure 2d.

Since $u$ has a better distinguishing ability than $v$ in both S1 and S2, the similarity was calculated by varying $u$ from 1 to 1000 while setting $v$ at 1 to simplify the calculation. Similarity results calculated by S1 and S2 are shown in Figure 3. As shown in Figure 3a1,a2, the similarity decreases linearly with...
increasing $u$ or peak height in both groups, indicating that S1 is more sensitive to a higher peak. However, the lower similarity in Figure 3a1,a2 of the same peak height represents higher sensitivity to a wider peak of S1. What is more, a sharper peak is easier to be distinguished as the results are displayed in Figure 3a3.

The similarity results of S2 for the three groups are shown in Figure 3b1−b3, which demonstrate that S2 is also more sensitive to a higher peak with the same peak width, the wider peak with the same peak height, and the sharper peak when the peak area is constant. The difference is that the results from S2 are much lower than S1, and this is also evidenced that S2 has a higher distinguishing ability than S1. The research showed that S1 and S2 could successfully distinguish one spectrum with another even small peak. Theoretically, S1 and S2 can distinguish extremely small peaks since the value of $u$ can reach positive infinity.

**Feasible Application for RMID with NIR.** **Spectral Investigation.** To demonstrate the feasibility of S1 and S2 algorithms for RMID, NIR spectra of five kinds raw materials including MCC PH101, MCC PH102, carboxymethyl starch (CMS), hydroxypropyl cellulose (HPC), and starch were used for analysis in this study. Figure 4a depicts raw spectra of five materials, showing broad peaks produced by combinations and overtones of hydrogen bonds (C−H, N−H, O−H, and S−H). The difference seen from the spectra makes identification for these materials possible. Spectral pretreatment using the first derivative could also amplify the effects of noise, and Savitzky−Golay smoothing (25 points) could effectively reduce the noise information. As displayed in Figure 4b, the baseline drift was minimized in the pretreated spectra and more resolvable peaks were observed.

**Sample Set Division.** To adapt and verify the accuracy of S1/S2, consideration should also be given to the sample set analyzed in each material. Spectra in the calibration set were utilized to calculate the similarity and define an appropriate range of factors, and those spectra in the validation set were shown the capability of S1/S2 in practical RMID. In this study, 30 samples were selected as a calibration set and the remaining 20 samples of each kind were used for method validation to get the average spectrum. The average spectrum of the calibration set was used as a reference spectrum, as referring to eqs 6 and 7.

**Selection of Optimal $u$.** To ensure adequate selectivity in establishing a classification model, one important parameter is to choose an appropriate threshold, which is the lowest value to correctly assign a given spectrum to a specific class. A low threshold could result in failing to distinguish different materials when their spectra are too similar. On the other hand, a high threshold can lead to failure to recognize its own materials belonging to the same category. The threshold in this study was set to 0.97, which is considered to be the material of the same category when the similarity is greater than 0.97 and judged to be different materials when the similarity is less than 0.97.

According to the previous results, all three factor selection methods ((i) varying $u$ with constant $v$, (ii) varying $v$ with constant $u$, and (iii) varying $u$ and $v$ simultaneously) demonstrate excellent distinguishing ability for PH101 and PH102. To simplify the parameter selection process, only varying $u$ from 1 to 50 with default $v$ was investigated. Spectra in the calibration set were calculated with the average spectra of PH101, PH102, HPC, CMS, and starch. The similarity of all spectra in the calibration set calculated with the average spectrum of PH101, PH102, HPC, CMS, and starch is depicted in Figure 5. As illustrated in Figure 5a, with increasing the value of $u$, the similarity of HPC, CMS, and starch decreases rapidly while PH102 decreases slowly. The difference in the similarity between the four kinds of raw materials with PH101 indicating the degree of difference is starch > CMS > HPC > PH102. The similarity results, as shown in Figure 5a, were oriented from maximum visibility of all of the pharmaceutical materials. HPC, CMS, and starch are well separated from PH101, and PH102 is partially separated at $u = 1$. It was of interest to know if the
analysis with S1/S2 would show unique characteristics of individuals; the distinct results presented in Figure 5 are consistent with the spectra difference in Figure 4. Considering the initial equation (eq 4) to deduce S1 and S2, if the lengths of the two compared spectra are similar, then their similarity (S1 and S2) value is mainly determined by the difference between their spectra, which is essentially attributed to the difference in the sample’s structure. For a given unknown spectrum, if its similarity with the reference spectrum is low and rejected according to the defined threshold, then it indicates that the spectral regions with large differences between them are mainly responsible for the level of similarity. These characteristic differences shown in the absorption of spectral regions are the foundation for us to extract chemical knowledge, which is closely related to chemical bonds. That is to say, the difference in spectral characteristic peaks owing to overtones and combination modes of functional groups containing hydrogen atoms, such as C−H, O−H, and N−H, is the foundation to discriminate different materials. On the other hand, the similarity calculated by S1 reveals a distinct separation of PH102 in the range of 4−14 of u, as shown in Figure 5b, which could be defined as the “available range”.

To define the available range for others, the similarity for all five kinds of materials was all calculated with S1 and S2. Under the condition that the threshold is 0.97, the values when the similarity to itself beginning to be lower than 0.97 and similarity to other materials all lower than 0.97 are summarized in Table 1. It can be seen from Table 1 that, when using S1 with variable u and constant v for the identification of these five raw materials, the similarity values of each kind of material to itself and others are different. The selection principle of u is to ensure that the similarity between raw materials and itself is higher than 0.97, and the similarity with other materials is lower than 0.97. For these five raw materials, it can be considered that the value of u between 4 and 14 could achieve accurate discrimination. When using S2 with variable u and constant v for the RMID, the similarity values of PH101, PH102, HPC, CMS, and starch to itself are lower than 0.97 when u is equal to 9, 11, 13, 8, and 6, respectively. Therefore, the range for the value of u of the S2 method to identify these five raw materials should be between 2 and 6.

Method Validation. To demonstrate the accuracy, S1, S2, the correlation coefficient, and cosine θ, PLS-DA, SVM, and SIMCA algorithms were all applied to the spectra of 20 batches for each material in validation sets. There is no parameter selection needed in the correlation coefficient and cosine θ. The algorithm of PLS-DA could be broadly regarded as a PLS regression between a descriptor matrix X and a response matrix Y, and SIMCA is another classical class-modeling tool incorporating principal component analysis (PCA), which are both focusing on searching for latent variables (LVs) with a maximum covariance. SVM is capable to select a hyperplane based on maximizing the distinguishing margin between classes and extend to nonlinear separation through the kernel machine scheme, while the appropriate kernel parameter choosing is the most critical. Generally, the implementation of S1/S2 could generate a variety of combinations of factors and the definition

---

**Table 1. Values of u when the Similarity of Materials to Itself Begins to be Lower than 0.97 and Similarity of Materials to Others is Lower than 0.97**

|       | PH101 | PH102 | HPC | CMS | starch |
|-------|-------|-------|-----|-----|--------|
| S1 to itself | 14    | 17    | 29  | 18  | 14     |
| to others    | 4     | 2     | 1   | 1   | 1      |
| S2 to itself | 9     | 11    | 13  | 8   | 6      |
| to others    | 2     | 2     | 1   | 1   | 1      |

---

Figure 4. (a) NIR spectra of MCC PH101, PH102, HPC, CMS, and starch and the (b) first derivative and Savitzky−Golay smoothing (25 points) of the spectra.

Figure 5. (a) Similarity of five materials and (b) PH101 and PH102 in the calibration set calculated with the average spectrum of PH101 of S1.
of useful $u/v$ and their optimization is essential to practical applications in RMID.

Strictly speaking, factor $u$ should be selected within a certain range that is greater than 4 and less than 14 for S1 and the minimum and maximum are 2 and 6 for S2, respectively, according to the results displayed in Table 1. It has to be aware that the minimum or maximum value in the range generated by the calibration process would run the risk of fail recognition or rejection; therefore, the median value in the interval was recommended to adopt to be well suited for class separation.

The sensitivity (%) and specificity (%) of different methods are shown in Table 2. The PLS-DA, SIMCA, and SVM results for pretreated spectra presented in Table 2 show 100% sensitivity and specificity, which indicates a clear separation of the samples from each validation set. However, the correlation coefficient and cosine results fail to reach 100% as these two methods are not able to identify PH101 from PH102 and distinguish between CMS and starch. The S1 and S2 results in Table 2 also displayed 100% sensitivity and specificity, showing a distinct division of all five kinds of materials. It is important to note that, unlike unsupervised classification methods, S1 and S2 are two class separation tools that need to be provided reference sample information. The sensitivity and specificity in this study was 100%, which is remarkably good for the spectra searching method.

To intuitively display the similarity results for the validation set acquired by S1 and S2, the data in Figure 6A,B were plotted. As shown in Figure 6A, the same materials can be detected and different materials can be rejected under the condition of $u=9$ and $v=1$ in S1. Similar results for S2 could be found in Figure 6B with the parameter $u=4$ when $v=1$. For the spectra searching method, it is excellent to achieve such a clear partitioning of different materials; the results of S1 and S2 demonstrate that both sensitivity and specificity are comparable to the classical supervised and unsupervised methods. Benefiting from the advantages of simple operation, intuitive information, and convenient library maintenance, both S1 and S2 methods could be the powerful alternative approaches for the RMID with NIR spectroscopy.

| methods          | CMS | HPC | PH101 | PH102 | starch |
|------------------|-----|-----|-------|-------|--------|
| PLS-DA           | 100 | 100 | 100   | 100   | 100    |
| SIMCA            | 100 | 100 | 100   | 100   | 100    |
| SVM              | 100 | 100 | 100   | 100   | 100    |
| correlation coefficient | 100 | 100 | 100   | 100   | 100    |
| cosine $\theta$  | 100 | 100 | 100   | 100   | 100    |
| S1 ($u=9$, $v=1$) | 100 | 100 | 100   | 100   | 100    |
| S2 ($u=4$, $v=1$) | 100 | 100 | 100   | 100   | 100    |
**APPLICATION STRATEGY**

To clarify the procedure in the application of RMID using S1 or S2, the steps could be summarized as follows:

a) By randomly selecting a certain number of sample spectra with known categories to construct a training set and a validation set, the first derivative with Savitzky–Golay smoothing was recommended to amplify the information and reduce noise interference.

b) According to the equations of S1 and S2, the average spectrum of each calibration set was taken as $X$ reference and all the spectra in calibration sets as $Y$.

c) Values of $u$ and $v$ distinguishing different categories could be optimized in different ways, such as varying $u$ and $v$ individually or simultaneously, and changing $u$ with $v$ set at default 1 was suggested in the calibration process.

d) An appropriate threshold, such as 0.97, was set to adjust the values of $u$ and $v$ to distinguish different categories.

e) The value of $u$ is usually a range, and the median value could be used to verify these two methods through the samples in the validation set.

f) If all samples can be successfully differentiated and identified, then the optimized $u$ and $v$ values are determined for the identification of unknown samples.

**CONCLUSIONS**

Two new similarity methods S1 and S2 that could both reflect the similarity from the perspective of the inner product of the two vectors related to vectorial lengths and the closeness with cosine $\theta$ or the correlation coefficient were proposed and investigated with NIR spectroscopy. It should be noted that S1 and S2 proposed in this study have viable and higher discrimination ability due to $u$ and $v$ factors ranging from 1 to $+\infty$. Calculation of $\Delta S1/\Delta u$, $\Delta S1/\Delta v$, $\Delta S2/\Delta u$, and $\Delta S2/\Delta v$ provided evidence that $u$ has a better discriminative ability than $v$. The similarity calculation based on small peaks in spectra indicated that the adjustment of $u$ and $v$ could distinguish any two spectra. During the comparison of the correlation coefficient and cosine $\theta$ for similarity evaluation of raw NIR spectra of five kinds of materials, S1 and S2 show obvious advantages compared to the correlation coefficient and cosine $\theta$. It should be mentioned that, as $u$ and $v$ factors can control the precision of the discrimination ability, it is the essential step to choose appropriate $u$ and $v$ to calculate similarities between the same and different materials.

The generally used algorithms for RMID are not always optimal for all applications. New similarity methods could 100% recognize the same material and distinguish different materials. Accordingly, this research could provide a simple and practical reference to establish similar methods for other pharmaceuticals. What is more, a combination of the advantages of several algorithms would yield very attractive and powerful tools and have a promising potential in pharmaceutical applications.

**MATERIALS AND METHODS**

**NIR Measurement.** The spectra of CMS, HPC, MCC PH101, PH102, and starch were acquired by using an Antaris II Fourier transform near-infrared spectrophotometer (Thermo Fisher Scientific, USA) in the wavenumber range of 10,000–4000 cm$^{-1}$ in the diffuse reflection mode with a 4 cm$^{-1}$ resolution. All spectra were averaged from 32 scans, and the gain was auto-optimized to increase the signal-to-noise ratio. To avoid the error brought from the environment, the temperature was constant at room temperature and humidity was at 30–50%.

**Data Analysis. Correlation Coefficient.** The correlation coefficient is a method for evaluating the similarity between spectra. Its equation is described as below

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

where $X_i$ represents the $i$th spectral intensity of the reference spectrum, and $\bar{X}$ is the mean value of the reference spectrum. $Y_i$ represents the $i$th spectral intensity of a sample’s spectrum, and $\bar{Y}$ is the mean value of the sample’s spectrum.

**Cosine $\theta$.** The cosine $\theta$ is the cosine of the angle between two compared vectors (i.e., spectra) for evaluating the similarity. Its equation was described as below

$$\cos \theta = \frac{\sum_{i=1}^{n}X_iY_i}{\sqrt{\sum_{i=1}^{n}X_i^2 \sum_{i=1}^{n}Y_i^2}}$$

in which $\theta$ is the angle between the two vectors $X_i$ and $Y_o$ representing the same ones in eq 3.

**Novel Similarity Methods.** The calculation formula is deduced in the following steps since

$$\frac{|X - Y|^2}{|X|^2 + |Y|^2} = \frac{X'X + YY' - 2XX'}{XX' + YY'} = 1 - \frac{2|X||Y| \cos \theta}{XX' + YY'}$$

where the symbol $'$ denotes the transpose operation and $||$ refers to the vector norm.

It can be seen from the formula (eq 3) that the initial mathematical meaning is that the numerator is the square of the difference between the two spectral vectors, and the denominator is the sum of the squares of the lengths of the two spectral vectors. The square term was used here for the convenience of data processing. If two spectral vectors are similar, then their difference is small; otherwise, the difference is large. However, if the similarity is evaluated by distance (i.e., their difference), then there is only the absolute quantity and no relative quantity (like vectorial angle cosine or the correlation coefficient), so it is difficult to understand the degree of similarity (or dissimilarity). For this reason, we have introduced the denominator part so that their difference can be measured in relative quantities.

According to vectorial operation, it is easy to deduce eq 4 from eq 3.

$$1 - \frac{|X - Y|^2}{|X|^2 + |Y|^2} = \frac{2|X||Y|}{XX' + YY'} \cos \theta$$

As shown in eq 4, the term $1 - \frac{|X - Y|^2}{|X|^2 + |Y|^2}$ should be not less than $-1$ and not more than $1$ ($-1 \leq 1 - \frac{|X - Y|^2}{|X|^2 + |Y|^2} \leq 1$) due to $0 \leq \frac{|X - Y|^2}{|X|^2 + |Y|^2}$ ≤ 2. If $X$ and $Y$ are in the same direction, then $X$ can be expressed as $X = kY$, and $\frac{|X - Y|^2}{|X|^2 + |Y|^2}$ can be simplified to

$$\frac{|X - Y|^2}{|X|^2 + |Y|^2} = 1 - \frac{2k}{1 + k^2}.$$
(0) can be acquired when \( k = -1 \) and 1, respectively. If \( X \) is orthogonal to \( Y \), then \( |X - Y|^2 = |X|^2 + |Y|^2 \) and \( 1 - \frac{|X - Y|^2}{|X|^2 + |Y|^2} = 0 \), which is also derived from the right part of eq 4. As seen from the term \( kX + Y\cos \theta \), its positive and negative value are determined by \( \cos \theta \). Since the properties of eq 4 conform to the common principle of general similarity, we will further expand eq 4 by the introduction of the sensitivity factor \( u \) and \( v \); see eq 5, and S1 (eq 6) has more adjustability and applicability so as to be suitable for handling complicated actual problems.

Then, when two parameters \( u \) and \( v \) are both positive numbers and greater than 1, as the product of \( u \) and \( v \) numbers and greater than 1, as the product of \( u \) and \( v \) tends to 0, we can obtain that

\[
\frac{2|X||Y|}{XX + YY} \cos \theta \geq \frac{2|X||Y|}{|X|^2 + |Y|^2} u (\cos \theta)^v \geq \frac{2|X||Y|}{|X|^2 + |Y|^2} u (\cos \theta)^v
\]

Therefore, our proposed similarity equation was defined as shown below (eq 6)

\[
S1 = \left( \frac{2|X||Y|}{|X|^2 + |Y|^2} \right)^u (\cos \theta)^v
\]

In eqs 5–8, \( X \) was the reference spectrum and \( Y \) was the spectrum of a sample to be determined. Exponents \( u \) and \( v \) are variable configurable parameters greater than 0 and less than infinity (\( 0 < u \) or \( v < +\infty \)) to accommodate practical applications. Generally, to increase the sensitivity of the S1 discriminating ability, \( u \) and \( v \) often take values greater than 1. Of course, to reduce the sensitivity, the values of \( u \) and \( v \) can also be adjusted toward 0. As shown in eq 4, we can see that

\[
\frac{|X - Y|^2}{|X|^2 + |Y|^2}
\]

reflects the ratio of the difference (squared term) of the two spectral vectors to the sum of the squares of their lengths. So, \( 1 - \frac{|X - Y|^2}{|X|^2 + |Y|^2} \) can indicate the similar degree of two spectral vectors. Based on eq 6, the first term of S1 reflects the similarity from the perspective of the inner product of the two spectral vectors, and the second term refers to the closeness with cosine of the vectorial angle. The sign of S1 is determined by cosine. The value of S1 was always greater than \(-1\) and less than 1. If \( X = kY \) where \( k \neq 1 \), then S1 can distinguish their difference in vectorial length by \( \frac{2|X||Y|}{|X|^2 + |Y|^2} \), but neither cosine \( \theta \) nor the correlation coefficient can do. When \( X = Y \), the S1 value will get its maximum of 1. What is more, when \( X = -Y \) with \( u \) and \( v \) set at 1 (i.e., their default values), the S1 value will get a minimum of \(-1\). If \( X \) is orthogonal to \( Y \), then S1 will be equal to 0.

In addition, if \( X \) and \( Y \) are both centralized by their means, then the similarity formula can also be derived as eq 7

\[
S2 = \left( \frac{2(|X - \bar{X}|)(|Y - \bar{Y}|)}{|X - \bar{X}|^2 + |Y - \bar{Y}|^2} \right)^u (\text{corr}(X, Y))^v
\]

In this equation, \( X \), \( Y \), \( u \), and \( v \) are same as the ones shown in eq 6. The difference is that \( X \) and \( Y \) are mean-centered, and cosine \( \theta \) is replaced with \( \text{corr}(XY) \) in eq 6 to get higher sensitivity, where \( \text{corr} \) represents the correlation coefficient.

**S1 and S2 Performance Evaluation.** Average NIR spectra of PH101 and PH102 were adopted to compare the \( u \) and \( v \) distinguishing ability with the rate of change \( \Delta S1/\Delta u \), \( \Delta S2/\Delta u \), \( \Delta S1/\Delta v \), and \( \Delta S2/\Delta v \). Based on the spectrum of PH101, three groups of small peaks including the wide peak group, narrow peak group, and constant peak area group were simulated with the Gaussian function to verify the sensitivity of S1 and S2 to a small peak.

**RMID Application.** The raw material data set, consisting of NIR spectra of 50 samples of MCC PH101, PH102, HPC, CMS, and starch, was collected in diffuse reflection mode. S1 and S2 were compared with classic spectra search methods, including the correlation coefficient and cosine \( \theta \), and supervised methods, such as PLS-DA, SVM, and SIMCA.

The sensitivity and specificity were both used to characterize the identification ability.

\[
\text{sensitivity} = \frac{N_r}{N_1} \times 100\%
\]

\[
\text{specificity} = \frac{N_{re}}{N_2} \times 100\%
\]

where \( N_r \) refers to the number of correctly recognized samples of its own category, and \( N_1 \) refers to the actual number of samples of its own category. \( N_{re} \) is the number of correctly refused samples of other categories, and \( N_2 \) indicates the actual number of samples of other categories.

**AUTHOR INFORMATION**

**Corresponding Authors**

Lei Nie — School of Pharmaceutical Sciences, Cheelee College of Medicine, Shandong University, Jinan, Shandong 250012, China; Phone: +86 531 88382330; Email: nieli2004@sdu.edu.cn

Hengchang Zang — School of Pharmaceutical Sciences, Cheelee College of Medicine, Shandong University, Jinan, Shandong 250012, China; orcid.org/0000-0003-2751-0281; Phone: +86 531 88380268; Email: zanghcw@126.com

**Authors**

Hui Zhang — School of Pharmaceutical Sciences, Cheelee College of Medicine, Shandong University, Jinan, Shandong 250012, China

Lian Li — School of Pharmaceutical Sciences, Cheelee College of Medicine, Shandong University, Jinan, Shandong 250012, China

Shuang Quan — School of Pharmaceutical Sciences, Cheelee College of Medicine, Shandong University, Jinan, Shandong 250012, China

Weilu Tian — School of Pharmaceutical Sciences, Cheelee College of Medicine, Shandong University, Jinan, Shandong 250012, China

Kefan Zhang — School of Pharmaceutical Sciences, Cheelee College of Medicine, Shandong University, Jinan, Shandong 250012, China

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.0c03831

**Notes**

The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**

We are grateful for the financial support of the National Key Research and Development Plan of China (2019YFC1711200).
and the Major Innovation Project of Shandong Province of China (2018CXSJC1405).

REFERENCES

(1) FDA Code of Federal Regulations 21 CFR 211.84. http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfCFR/CFRSearch.cfm?fr=211.84.

(2) Good Manufacturing Practice Guide for Active Pharmaceutical Ingredients, Draft Consensus Guideline, International Conference on Harmonisation (ICH), 2000.

(3) Blanco, M.; Romero, M. A. Near-infrared libraries in the pharmaceutical industry: a solution for identity confirmation. Analyst 2001, 126, 2212–2217.

(4) Sun, L.; Hsiung, C.; Pederson, C. G.; Zou, P.; Smith, V.; von Gunten, M.; O’Brien, N. A. Pharmaceutical Raw Material Identification Using Miniature Near-Infrared (MicroNIR) Spectroscopy and Suppressed Pattern Recognition Using Support Vector Machine. Appl. Spectrosc. 2016, 70, 816–825.

(5) Kumar, M.; Bhatia, R.; Rawal, R. K. Applications of various analytical techniques in quality control of pharmaceutical excipients. J. Pharm. Biomed. Anal. 2018, 157, 122–136.

(6) Hespanhol, M. C.; Souza, J. C.; Pasquini, C. Feasibility of a portable, low-cost near-infrared spectrophotometer for the quality screening of omega-3 dietary supplements. J. Pharm. Biomed. Anal. 2020, 189, 113436.

(7) Calvo, N. L.; Alvarez, V. A.; Lamas, M. C.; Leonardi, D. New approaches to identification and characterization of tioconazole in raw material and in pharmaceutical dosage forms. J. Pharm. Anal. 2019, 9, 40–48.

(8) Tie, Y.; Duchateau, C.; Van de Steene, S.; Mees, C.; De Braekeleer, K.; De Beer, T.; Adams, E.; Deconinck, E. Spectroscopic techniques combined with chemometrics for fast on-site characterization of suspected illegal antimicrobials. Talanta 2020, 217, 121026.

(9) Dong, Qi; Yu, Z.; Li, L.; Nie, L.; Li, D.; Zang, H. Near-infrared spectroscopic study of molecular interaction in ethanol-water mixtures. Spectrochim. Acta, Part A 2019, 222, 117183.

(10) Near-Infrared Spectrophotometry; United States Pharmacopoeia, USP, 2015, 38, NF33.

(11) Near-Infrared Spectroscopy, European Pharmacopoeia, 8th ed. 2013.

(12) Omata, R.; Hattori, Y.; Sasaki, T.; Sakamoto, T.; Otsuka, M. Elucidation of the Molecular Mechanism of Wet Granulation for Pharmaceutical Standard Formulations in a High-Speed Shear Mixer Using Near-Infrared Spectroscopy. Pharmaceuticals (Basel) 2020, DOI: 10.3390/ph13090226.

(13) Pawar, P.; Clancy, D.; Gorringer, L.; Barlow, S.; Hesketh, A.; Elkes, R. Development and Scale-Up of Diversion Strategy for Twin Screw Granulation in Continuous Manufacturing. J. Pharm. Sci. 2020, 3439.

(14) Avila, C. R.; Ferré, J.; de Oliveira, R. R.; de Juan, A.; Sinclair, W. E.; Mahdi, F. M.; Hassanpour, A.; Hunter, T. N.; Bourne, R. A.; Muller, F. L. Process Monitoring of Moisture Content and Mass Transfer Rate in a Fluidised Bed with a Low Cost Inline MEMS NIR Sensor. Pharm. Res. 2020, 37, 84.

(15) Domokos, A.; Nagy, B.; Gyüzéki, M.; Farkas, A.; Tacsí, K.; Patakai, H.; Liu, Y. C.; Balogh, A.; Firth, P.; Szilágyi, B.; Marosi, G.; Nagy, Z. K.; Nagy, Z. K. End-to-end continuous manufacturing of conventional compressed tablets: From flow synthesis to tabling through integrated crystallization and filtration. Int. J. Pharm. 2020, 581, 119297.

(16) Chu, X.-L.; Li, J.-Y.; Chen, P.; Xu, Y.-P. Algorithms, Strategies and Application Progress of Spectral Searching Methods. Chin. J. Anal. Chem. 2014, 42, 1379–1386.

(17) Batchellor, N. Cluster-Analysis. Chemom. Intell. Lab. Syst. 1989, 6, 105–125.

(18) Ferreiro-González, M.; Espada-Bellido, E.; Guíllén-Cueto, L.; Palma, M.; Barroso, C. G.; Barbero, G. F. Rapid quantification of honey adulteration by visible-near infrared spectroscopy combined with chemometrics. Talanta 2018, 188, 288–292.