Current field-based chemical identification instruments for pharmaceutical applications typically use one of three analytical methodologies: hit quality index (HQI), traditional chemometrics, or the Thermo Scientific probabilistic approach. This white paper compares HQI and the Thermo Scientific probabilistic approach, leaving the comparison between probability and traditional chemometrics to a separate white paper.

**Hit Quality Index**

Traditional methods for reference-library searching are typically based on the assessment of similarity metrics calculated via peak table comparisons, or more commonly, from those generated by full spectrum comparisons. Full spectrum approaches typically generate a “hit quality index” (HQI) between the unknown spectrum and each library spectrum. The HQI can be calculated based on Euclidean distance, median absolute deviation, or perhaps most frequently, the correlation coefficient between the test spectrum and each library spectrum. The correlation coefficient is equivalent to measuring the cosine of the angle between two spectra. The resulting correlation coefficient, $R$, is 1 when the two spectra are in perfect correspondence and 0 when they are orthogonal.

While a correlation coefficient threshold of 0.95 is frequently used to determine whether two spectra are a match, the correlation is merely an angle and not a probability. Thus, the traditional threshold of 0.95 in no way means 95% likelihood, 95% confidence, or 95% agreement. Furthermore, a correlation coefficient other than 0 or 1 has no direct interpretation in the context of spectral identity because a transparent interpretation as a test statistic only holds when dealing with random normal variates, clearly not the case for Infrared or Raman spectra. While the correlation coefficient has been a popular choice for pure material assessment, it is not particularly sensitive to discrepancies between spectra of interest.

**Probabilistic Evaluation**

As technical advances brought laboratory-quality instruments to the field, a new testing approach was needed to address the challenge of unknown chemical identification. In the process of identifying substances within a vast unknown library, handheld instruments put the power of spectroscopy into the hands of a new user – field technicians without extensive spectroscopy and chemical training. While HQI met the initial need for laboratory use, a new approach was required for these less experienced users who operate in challenging environments and sampling conditions.
Table 1. In an evaluation of celluloses, the p-value approach correctly identifies the sample as different from the library reference (Microcrystalline Cellulose) while HQI does not.

<table>
<thead>
<tr>
<th>Cellulose</th>
<th>p-value</th>
<th>HQI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microcrystalline Cellulose</td>
<td>0.338</td>
<td>0.9998</td>
</tr>
<tr>
<td>Hydroxyethyl Cellulose</td>
<td>0.00000754</td>
<td>0.9970</td>
</tr>
<tr>
<td>Methyl Cellulose</td>
<td>0.00000185</td>
<td>0.9766</td>
</tr>
<tr>
<td>Hydroxypropyl Cellulose</td>
<td>0.0000000323</td>
<td>0.9796</td>
</tr>
</tbody>
</table>

As values in Table 1 illustrate, the probabilistic and HQI approaches both correctly identify Microcrystalline Cellulose. In the probabilistic approach, we accept the null hypothesis (p-value > 0.05) and, in the HQI approach, the correlation coefficient is very nearly 1.0. However, in regard to the other celluloses, the probabilistic approach rejects the null hypothesis (p-value < 0.05) while the correlation method suggests reference matches with these materials – clearly returning false-positive results.
In Figure 1, we can visually examine the measured spectrum of Hydroxyethyl Cellulose (the reference sample) versus the Microcrystalline Cellulose library spectrum. The p-value for the Hydroxyethyl Cellulose sample spectrum, shown in Table 1, is 0.00000754. The p-value result below 0.05 indicates a low probability that the measured spectrum of the unknown sample lies within the multivariate domain of a reference spectrum of interest, if the null hypothesis is true (i.e. the unknown and library are different). Meanwhile, the HQI result of 0.9970 is a high correlation score (e.g. cosine angle), far above the traditional 0.95 passing threshold, yet the unknown material is Hydroxyethyl Cellulose, not Microcrystalline Cellulose.

To further illustrate the effectiveness of the probabilistic approach, we will consider the probabilistic comparison of 15% DMMP in chloroform to pure chloroform, as shown in Figure 2. In this case, we examine the unknown measured spectrum and the pure chloroform library spectrum, testing the null hypothesis (H₀ = pure chloroform) and the alternative (H₁ = not pure chloroform). With this assessment, it becomes very clear that the discrepancy in the 715 cm⁻¹ region cannot be due to noise alone. The p-value is the probability of observing the unknown spectrum or one more extreme, if the null hypothesis is true. For this comparison, the value is calculated as 6.1 x 10⁻⁴. Thus, if the sample were pure chloroform, the probability of observing a spectrum as extreme as the unknown measurement would be ~ 1 in 1639 – highly unlikely. Correspondingly, the algorithm would recognize that the sample cannot be pure chloroform, returning a p-value less than 0.05 (i.e. statistically significant).
Summary
Both HQI and probabilistic methods are proven analytical techniques for interpretation of spectroscopic data. While HQI is well-suited for laboratory use by spectroscopy experts – its original and intended purpose – probabilistic analysis is specifically designed for field-based decision making, with very high accuracy. When considering these options, users should evaluate the simplicity and reliability of results in relation to relatively inexperienced users who operate in challenging environments and sampling conditions.

References
2. TD Wickens, Elementary Signal Detection Theory, Oxford University Press 2001
3. GM Pesyna et al. “Probability based matching system using a large collection of reference mass spectra”
6. CD Brown and RL Green, “Considerations for embedded authentication using intelligent portable Raman systems,”
   FACSS 2006, Orlando, FL