KnowItAll Software Training

SmartSpectra Databases and Classification Models

WILEY

Using SmartSpectra in KnowItAll

Purpose

These exercises demonstrate how to use SmartSpectra in KnowItAll ID Expert and SearchIt.

Objectives

These exercises will teach you:

- How to use SmartSpectra IR spectra in KnowItAll SearchIt
- How to use SmartSpectra Raman spectra in KnowItAll SearchIt
- How to use SmartSpectra IR spectra in KnowItAll ID Expert
- How to use SmartSpectra Raman spectra in KnowItAll ID Expert

Training Files Used in This Lesson:

- 4-(Pyridin-3-yl)-2-2,6,2-terpyridine.irf
- 4,13-DIDECYL-1,7,10,16-TETRAOXA-4,13-DIAZACYCLOOCTADECANE.irf
- X-Phos.irf
- 1-acetyl-1234-tetrahydroquinoline.irf
- 17-Hydroxy-17-alpha-pregn-4-en-20-yn-3-one.irf

Note: The training files used are for example purposes only. The user should utilize their own IR spectra when following this training.

KnowItAll Applications Used:

- KnowItAll SearchIt
- KnowItAll ID Expert

Background

Wiley has the largest quantity of experimental IR spectra, which includes the Sadtler IR collection. Even with the continued development in the chemical industry, there is still insufficient progress to cover the newly-discovered chemical space. Sample collection is also challenging, time-consuming and costly. Wiley's IR SmartSpectra Collection is an attempt to increase the catalog coverage of Wiley's IR chemical space. This does not mean that this database will increase the chemical space, but that it will increase the amount of coverage within the bounds of our current libraries' chemical space. With the development of computer modeling technology, the feasibility of using computed IR spectra to compensate for the lack of samples was investigated and found to increase unknown compound classification.

These libraries can be helpful when used as tools for searching unknown compounds, illuminating the composition of an unknown spectrum, or accurately predicting the compound's spectrum for the associated structure and functional groups. Due to the predicted nature of this data, it should be mentioned that the results have a margin of error. Wiley acknowledges that some results may not be entirely correct. It is advised that users employ these libraries as additional tools to help the user classify components that are in the spectrum in order to characterize their unknown spectrum.

KnowItAll IR and Raman Search Algorithms

A background in the algorithms used by KnowItAll will be beneficial. For IR and Raman spectral comparison, KnowItAll uses the following algorithms:

Correlation

This is the default algorithm for searching in KnowItAll and it conforms to the industry standard for correlation algorithms. The Correlation algorithm is similar to the Euclidean Distance algorithm. The difference between the two is in the way the spectra are treated before the comparison. Each spectrum is mean-centered prior to performing the dot product normalization. This approach can improve search results for noisy spectra and spectra that have baseline issues, particularly with a baseline offset that is the result of a negative spike or chemical noise. It is slightly more time-consuming than the Euclidean Distance algorithm. The search speed is slower because each spectrum in the database must be mean-centered and then normalized prior to the comparison. The search results that are obtained with the Correlation algorithm are spectrally similar to the unknown, even if the unknown compound is not in the database. The Correlation algorithm is heavily weighted by peak areas. Broad features are weighted much more strongly than sharp features. The algorithm is most tolerant to peak shifts and non-linearities in relative band intensities.

Correlation (Classic)

The Correlation algorithm that was found in all versions of KnowItAll prior to KnowItAll 2020 is similar to the Euclidean Distance algorithm. However, it did not conform to the industry standard for correlation algorithms. Beginning with KnowItAll 2020, the Correlation algorithm does conform to the industry standard and it is the default algorithm used for searching in KnowItAll. To provide backward compatibility for customers who want to reproduce prior search results, the previous correlation algorithm is now provided as Correlation (Classic).

Euclidean Distance

The Euclidean Distance algorithm measures the point-to-point differences between a pair of spectra. The results that are obtained with the Euclidean Distance algorithm are spectrally similar to the unknown, even if the unknown compound is not in the database. This algorithm, however, can yield degraded search results when the unknown spectrum has a sloping or offset baseline. The Euclidean Distance algorithm is heavily weighted by peak areas. Broad features are weighted much more strongly than sharp features. This algorithm is most tolerant to peak shifts and non-linearities in relative band intensities.

First Derivative Euclidean Distance
Use this algorithm to reduce the effects of baseline slope or offset in the unknown. Although the search speed is slightly slower than with the Euclidean Distance algorithm, the First Derivative Euclidean Distance sometimes gives improved search results, especially when the unknown spectrum is a mixture of two or more compounds. The First Derivative Euclidean Distance algorithm is heavily weighted by changes in slope. Sharp features are weighted much more strongly than broad features. The algorithm is also very sensitive to peak shifts. Small shifts can make the algorithm miss a similar result.

Second Derivative Euclidean Distance

Use the Second Derivative Euclidean Distance algorithm to compare the second derivative of a reference spectrum to that of the query spectrum.

Optimized Corrections: A Breakthrough Technology for Spectral Searching

Spectral searching is one of the most important tools researchers use to classify and/or identify materials, yet it continues to be plagued by errors and imperfections. During a spectral search, a sample spectrum is compared to a database of reference spectra. To ensure that an optimal match is found in the database, spectra can be adjusted to compensate for differences between spectra caused by variability in instruments, accessories, environmental conditions, and other factors.

According to ASTM's guide on spectral searching[1], various algorithms and manual methods exist to adjust spectra to get reasonable match scores when two compared spectra of the same compound differ for various reasons. While these methods may work in selected cases, subtle discrepancies such as a shift of the X-axis are very hard to identify and correct manually. The inflexible mathematical algorithms typically employed do not compensate for these types of errors in spectra that are flawed.

Manual corrections can be made by expert spectroscopists, but those less experienced in spectroscopy are often unaware of how to perform the necessary corrections on their sample spectrum to achieve the best search result. To address this growing concern, Wiley has introduced Optimized Corrections, a breakthrough patented technology that performs a computationally complex set of multiple corrections on query and reference spectra in a search to find the optimal match between the query and each individual reference spectrum. This training guide will demonstrate how the Optimized Corrections technology yields better matches between query and reference spectra than can be attained using rigid search algorithms alone or with manual methods to optimize spectra for searching.

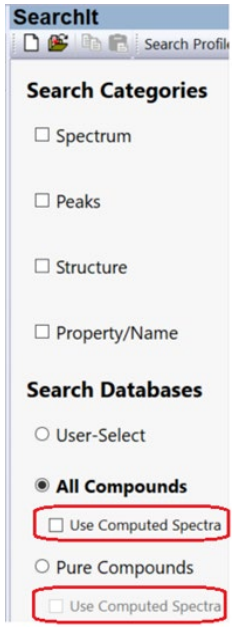
Optimized Corrections consider the full spectrum during a selected range(s) search.

[1] E2310-04 - Standard Guide for Use of Spectral Searching by Curve Matching Algorithms with Data Recorded Using Mid-Infrared Spectroscopy, 2009. ASTM International Web Site. http://www.astm.org/Standards/E2310.htm (accessed March 4, 2015).

SearchIt

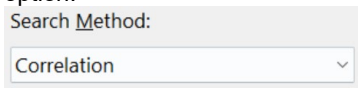
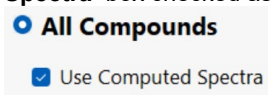
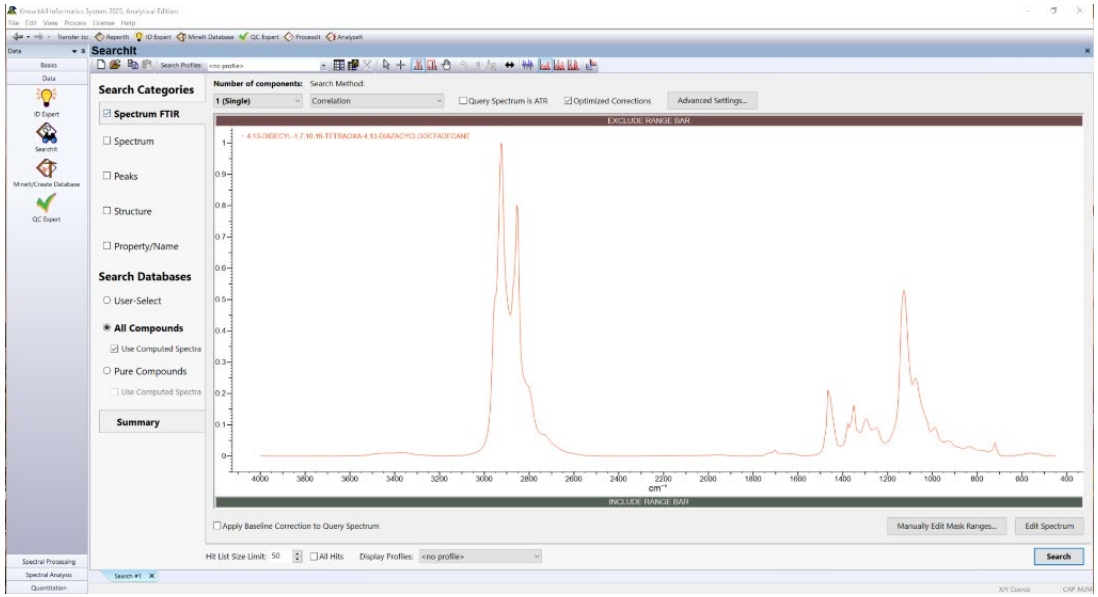
Introduction

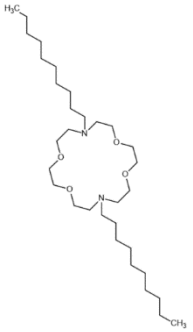
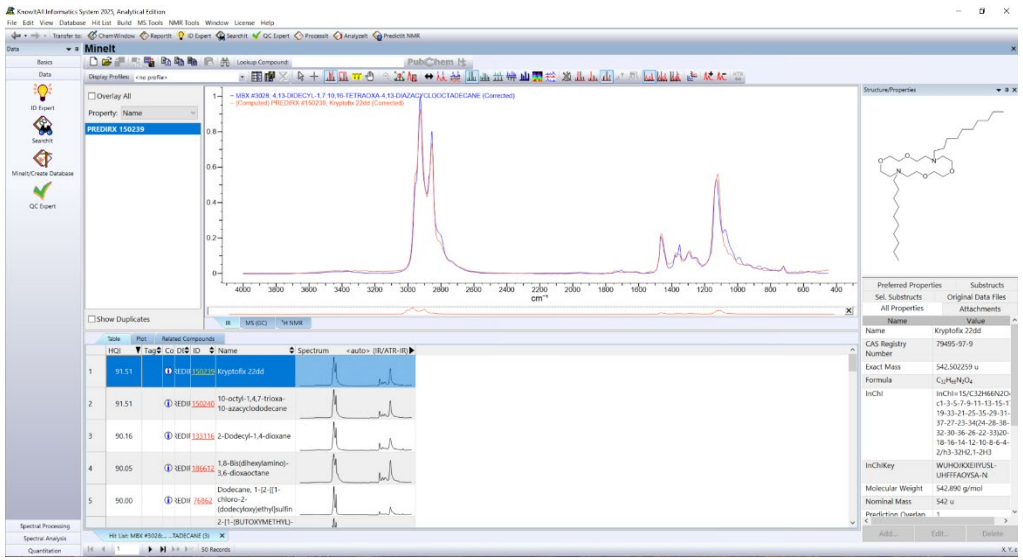
When a user license contains the subscription to SmartSpectra IR spectra, it can be accessed in the **SearchIt** application using **Search Databases > All Compounds** or **Databases > Pure Compounds** and by checking the box for the **Use Computed Spectra** option:



Example 1

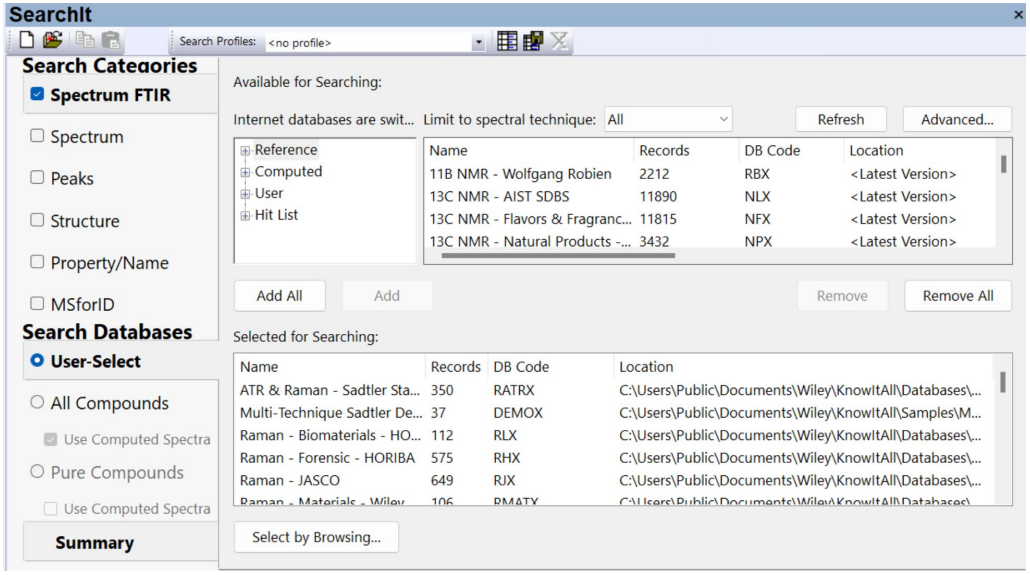
Example File: 4,13-DIDECYL-1,7,10,16-TETRAOXA-4,13-DIAZACYCLOOCTADECANE.irf

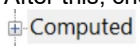
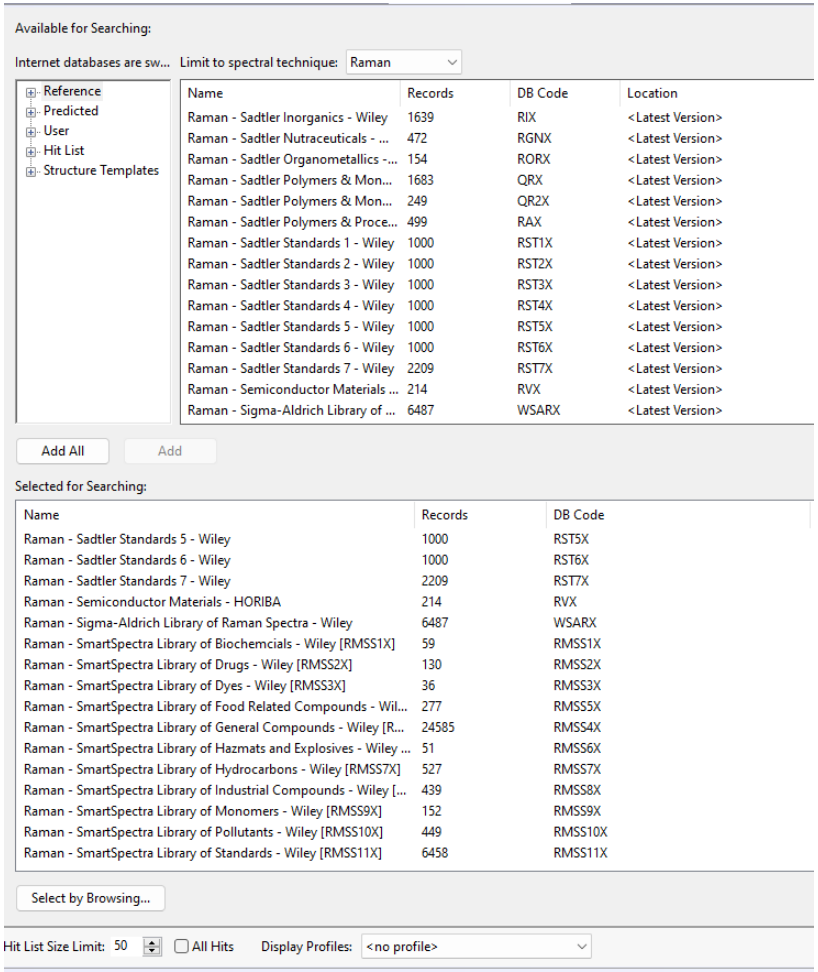
	Action	Result
1	In the **SearchIt** application, click **Open Spectrum or Structure** icon ( ) located on the **Standard Toolbar**.	*Note*: The training files used in this document are for example purposes only. The user should utilize their own IR spectra when following this training.
2	Navigate to find **4,13-DIDECYL-1,7,10,16-TETRAOXA-4,13-DIAZACYCLOOCTADECANE.irf** in **"C:\Users\Public\Documents\Wiley\KnowItAll\Samples\Computed Spectra\IR"** folder. Click **Open**. Confirm that the **Search Method** is set to **Correlation** algorithm. To change the method, use the dropdown menu to select it as the option.  Under the **Search Databases** section, select **All Compounds** with the **'Use Computed Spectra'** box checked as well.  Click **Search**.	The image on this page is the correct compound structure associated with this spectrum. Upon executing the search, **MineIt** application displays the query results. 

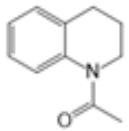
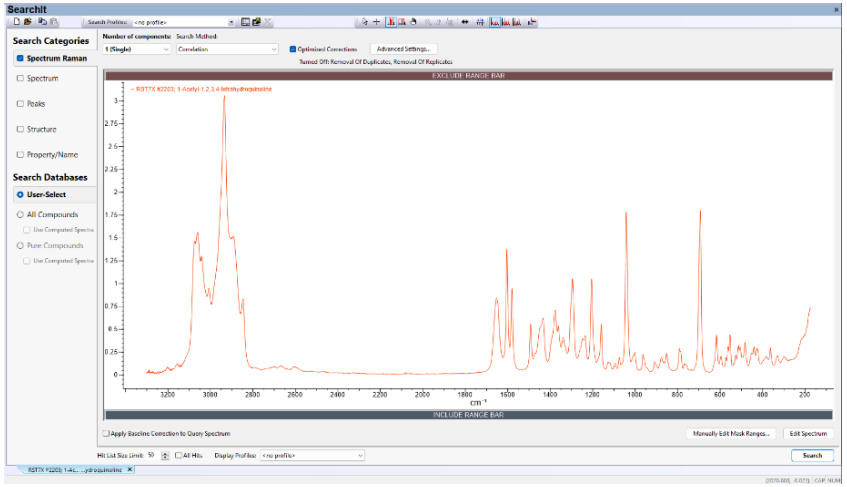
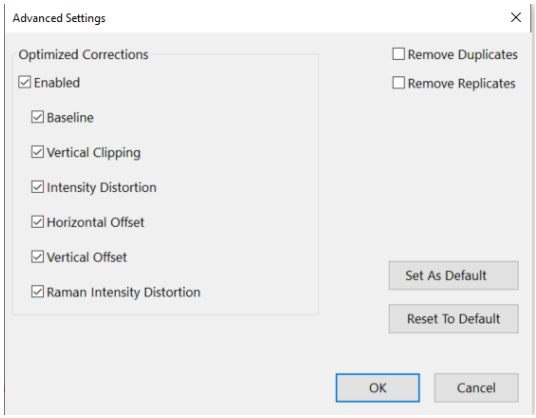
	Action	Result
3	In **MineIt**, the hit list table displays match results featuring a mix of predicted and empirical matches. The structure below is the correct match for the query spectrum's structure,. The structure below is an exact match with the predicted data which appears as the top hit in the table. 	The hit list results open in **MineIt**. The best match for the query appears at the top of the hit list. This top match is automatically selected and displayed:  *Note:* Specific results will depend on the available databases in the user license.

Example 2

Example File: 1-acetyl-1234-tetrahydroquinoline.irf

	Action	Result
1	Return to the **SearchIt** application. Clear the previous search be hitting the small **X** on the upper right-hand side of the screen (✕). In the **SearchIt** application, click **User-Select** under the **Search Databases** section.	The **User-Select Databases** window is displayed: 

	Action	Result
2	Using the **Limit to Spectral Technique** dropdown menu, choose **Raman**.  Click **Reference** () and then click **Add All** to add all licensed databases. After this, choose the **Computed** tab () and then click **Add All** to add all SmartSpectra databases.	After selecting and adding the **References** and **Computed** databases, all licensed Raman libraries will be available in the **Selected for Searching** window: 

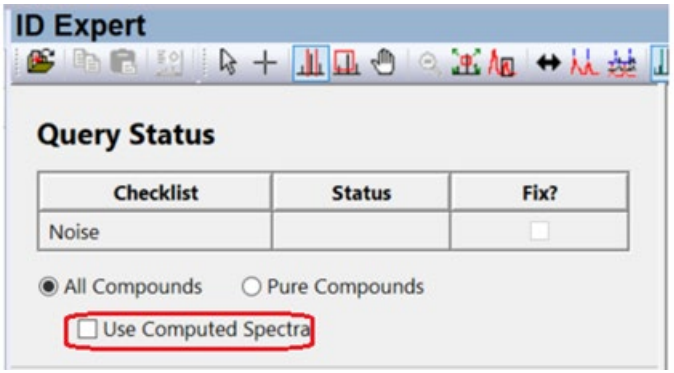
	Action	Result
3	Click **Open Spectrum or Structure** icon ( ) located on the **Standard Toolbar**. Navigate to find **1-acetyl-1234-tetrahydroquinoline.irf** in **"C:\Users\Public\Documents\Wiley\ KnowItAll\Samples\ Computed Spectra\Raman**". Click **Open**. *Note:* The image on this page is the correct compound structure associated with this spectrum. 	The selected file opens in the **Spectrum Search** window: 
4	In the **Spectrum Search** window, click **Advanced Settings** and deselect **Remove Duplicates** and **Remove Replicates**. Click **OK**. Click **Search**.	The **Advanced Settings** pop-up window is displayed. Upon clicking **OK** the window closes. Upon executing the search, **MineIt** loads with the query results. 

	Action	Result
5	Locate the **Hit List** to view match results featuring a mix of predicted and empirical matches.	The top hit in the **Hit List** is the empirical spectrum that best matches the query spectrum. The second best hit is from the SmartSpectra database. *Note:* In this simulated test, the top hit is to be disregarded as it was used as the query spectrum, thus the SmartSpectra result is the true best match. 

ID Expert

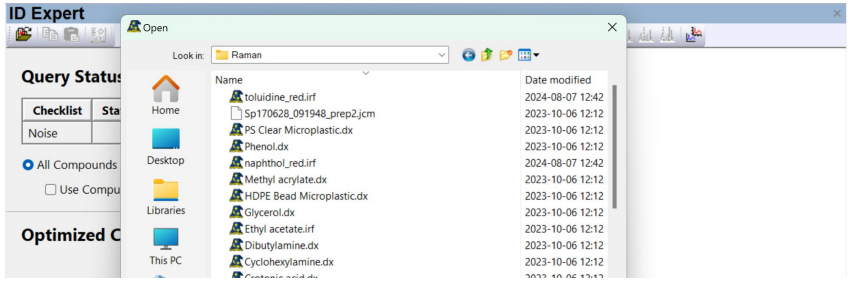
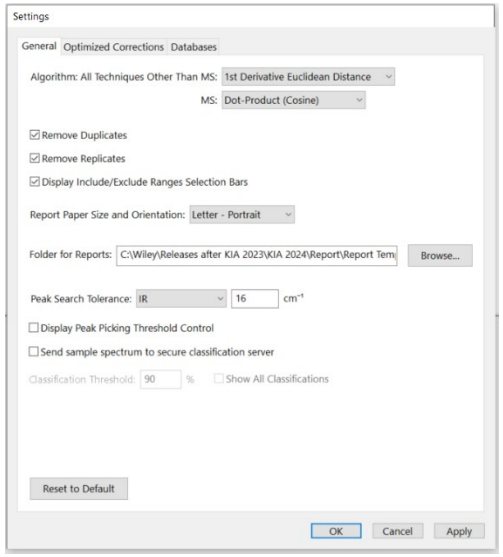
Introduction

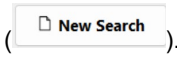
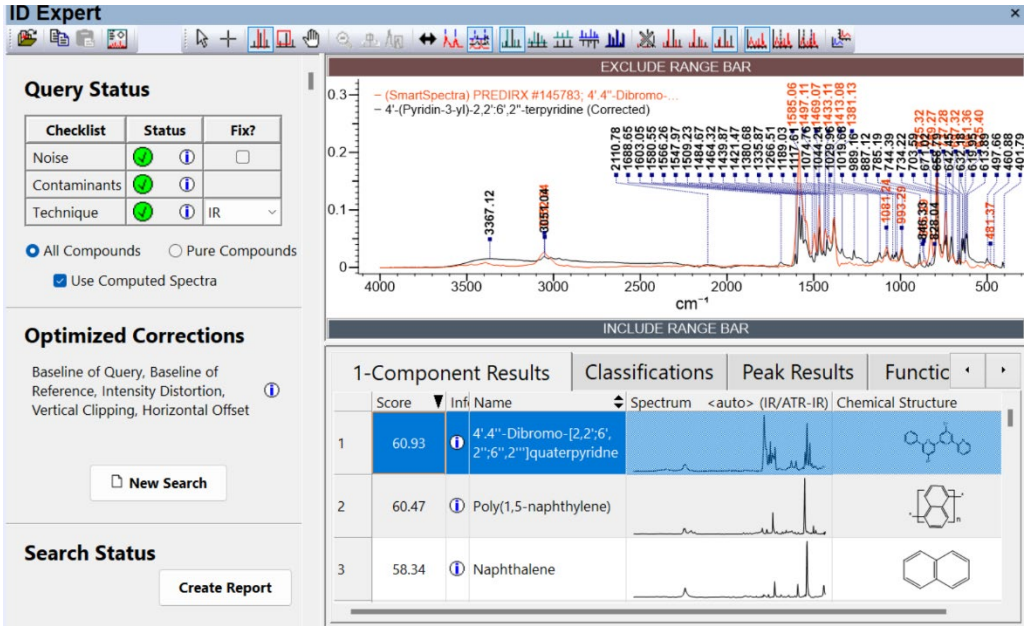
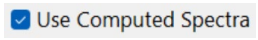
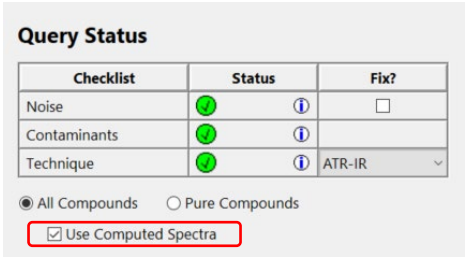
When a user's license contains the subscription for predicted IR spectra, the **ID Expert** interface **Query Status** shows the **Use Computed Spectra** option:

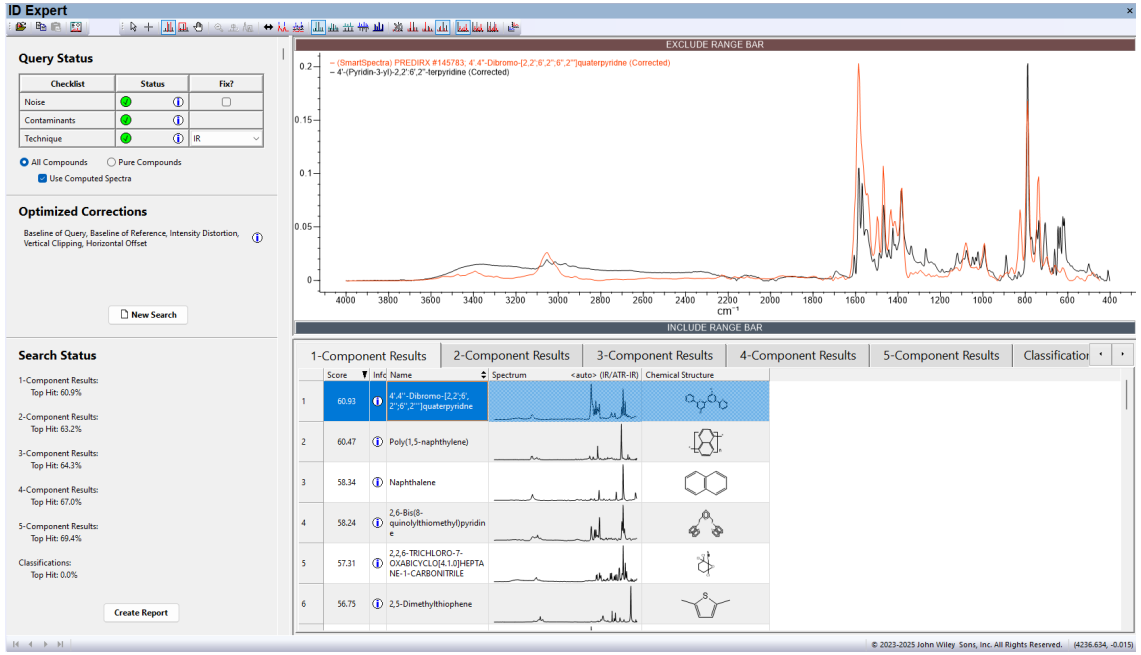


Example 3

Example File: 4-(Pyridin-3-yl)-2-2,6,2-terpyridine.irf

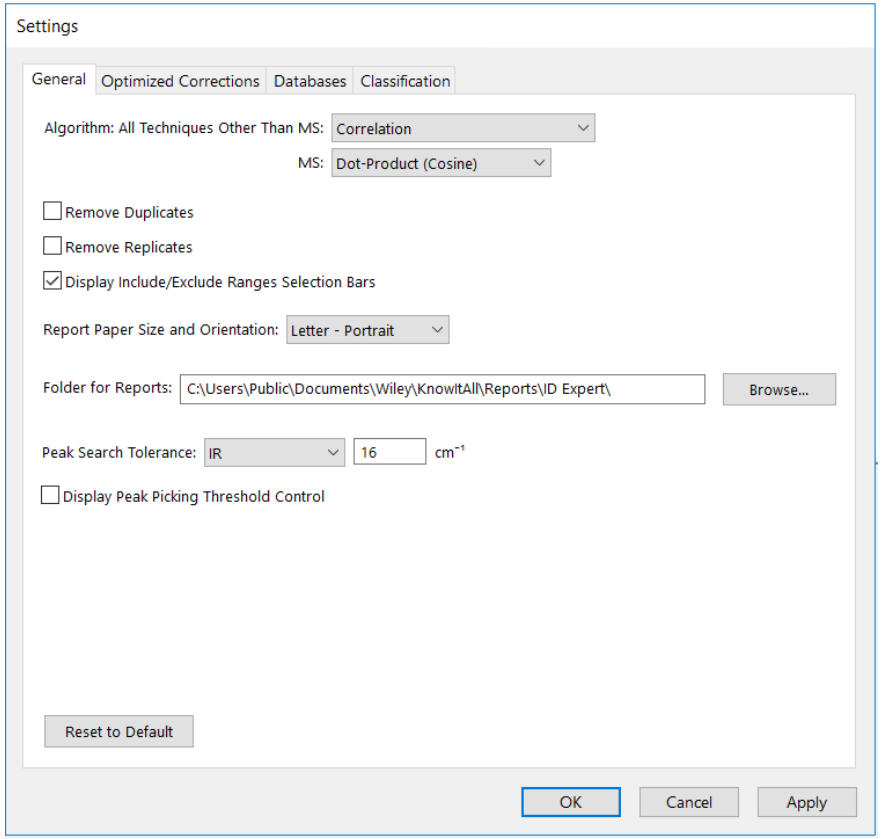
	Action	Result
1	Navigate to the **Data** toolbox and open the **ID Expert** application by clicking the **ID Expert** icon (ID Expert). Alternatively, if the desktop (standalone) application is installed, **ID Expert** can be opened directly by double-clicking on the desktop icon.	The application opens and an **Open** window displays: 
2	Click **Cancel** on the **Open** dialog window, then choose **File > Settings**. Under **General** tab, set **Algorithm: All Techniques Other Than MS** to menu option **1st Derivative Euclidean Distance** using the dropdown menu. Select **Apply** then **OK** to apply the change.	Upon clicking **Settings**, the **Settings** window opens: 

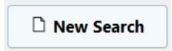
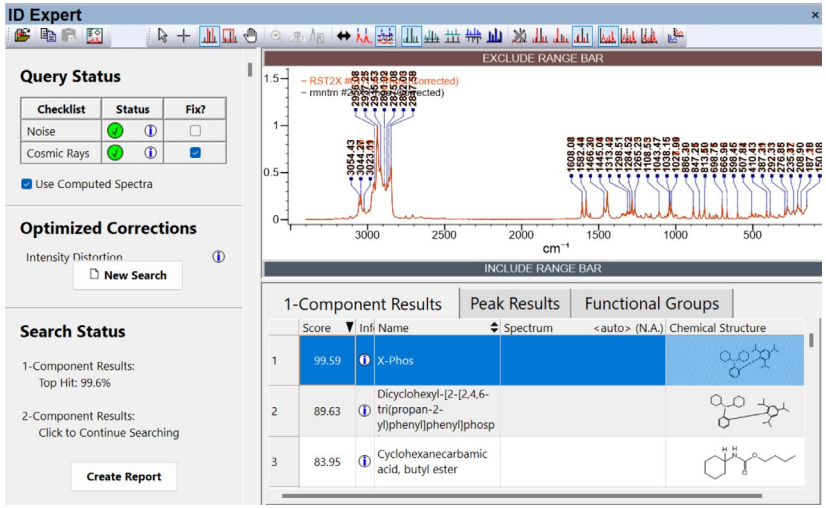
	Action	Result
3	Click the **New Search** button (). Open **4-(Pyridin-3-yl)-2-2,6,2-terpyridine.irf** in the **"C:\Users\Public\Documents\Wiley\KnowItAll\Samples\Computed Spectra\IR"** folder. Click **Open**. (*Note*: The training files used in this document are for example purposes only. The user should utilize their own IR spectra when following this training.)	Immediately upon opening the file, it begins to search for a match using the licensed databases available to the user: 
4	Under **Query Status**, change **Technique** to **ATR-IR**. Check **Use Computed Spectra** checkbox. 	The modified **Query Status** settings are displayed. **ID Expert** updates the search with the modified settings. 

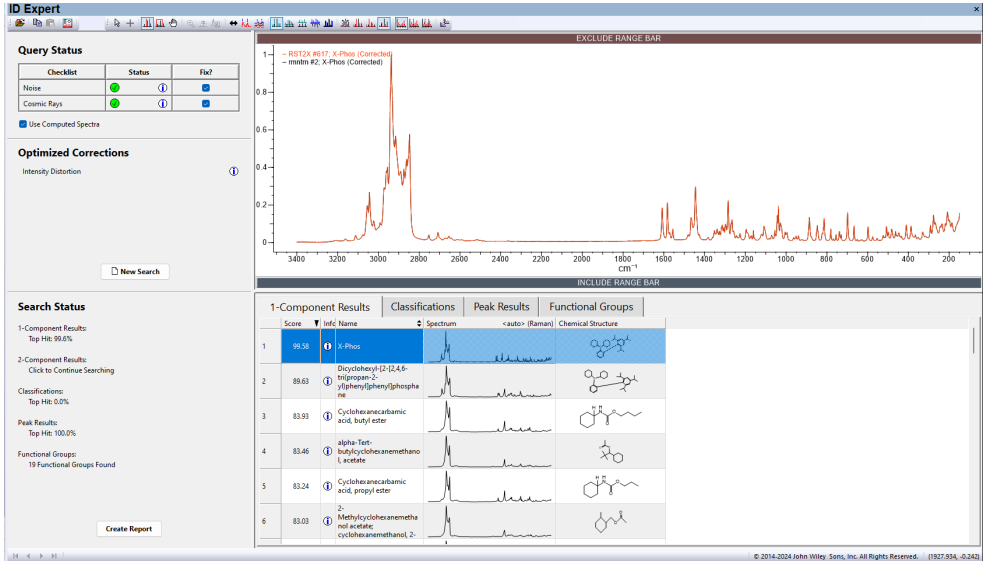
	Action	Result
5	View the search results in the **Component Table**.	This search results include experimental and predicted spectra. 
6	Close the active search by selecting the **X** icon ( ) in the top right-hand corner.	

Example 4

Example File: X-Phos.irf

	Action	Result
1	Close the **Open** window (if open), then choose **File > Settings**. In the **Settings** pop-up window, under the **General** tab, set **Algorithm: All Techniques Other Than MS** to menu option **Correlation** using the dropdown menu. Uncheck **Remove Duplicates** and **Remove Replicates**. Click **Apply** the **OK** to apply the change.	

	Action	Result
2	Click the **New Search** button (). Open **X-Phos.irf** in **"C:\Users\Public\Documents\Wiley\KnowItAll\Samples\Computed Spectra\Raman"**. Click **Open.** In the **Query Status** panel, check the **Use Computer Spectra** box. Checking the **Use Computer Spectra** checkbox adds SmartSpectra data to **ID Expert's** library search. When not checked, SmartSpectra will not be searched. (*Note*: The training files used in this document are for example purposes only. The user should utilize their own Raman spectra when following this training.)	Immediately upon opening the file, it begins to search for a match using the licensed databases available to the user: 

	Action	Result
3	View search results in the **Component Results** table.	The top hit in the **Hit List** is the empirical spectrum that best matches the query spectrum. The second best hit is from the SmartSpectra database. *Note:* In this simulated test, the top hit is to be disregarded as it was used as the query spectrum, thus the SmartSpectra result is the true best match. 
4	Close the active search by selecting the **X** icon ( ) in the top right-hand corner.	

Classification Models

Purpose

These exercises demonstrate how to use classification models in KnowItAll ID Expert and MineIt.

Objectives

These exercises will teach you:

- How to use the KnowItAll Classification models in ID Expert
- How to use the KnowItAll Classification models in MineIt

Training Files Used in This Lesson:

- 17-Hydroxy-17-alpha-pregn-4-en-20-yn-3-one.irf
- SmartSpectraFTIROxycodone.irf
- SmartSpectraRamanAndrosteroneacetate.irf

Note: The training files used are for example purposes only. The user should utilize their own IR spectra when following this training.

KnowItAll Applications Used:

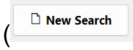
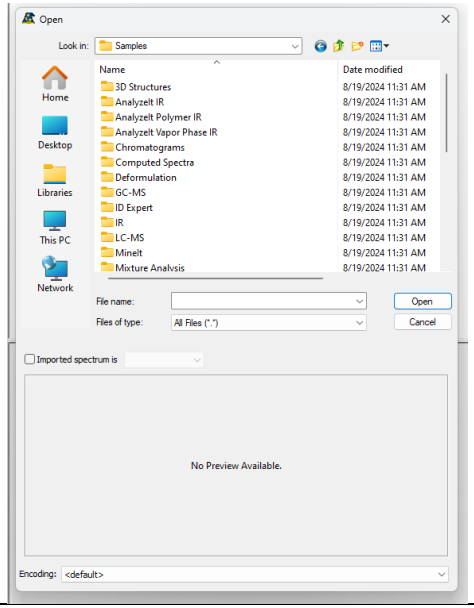
- KnowItAll ID Expert
- KnowItAll MineIt

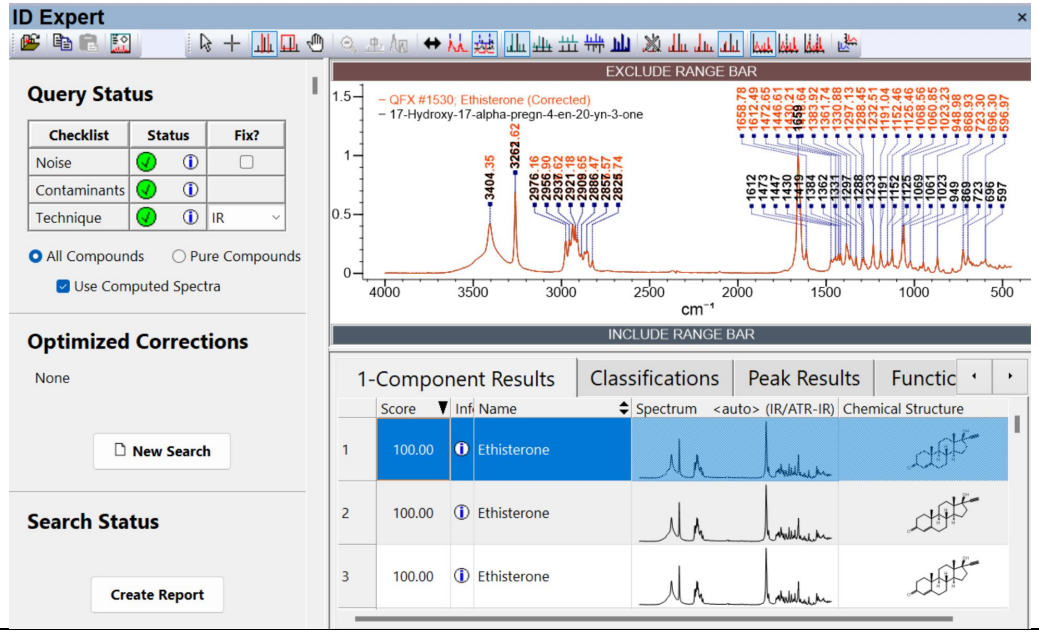
Background

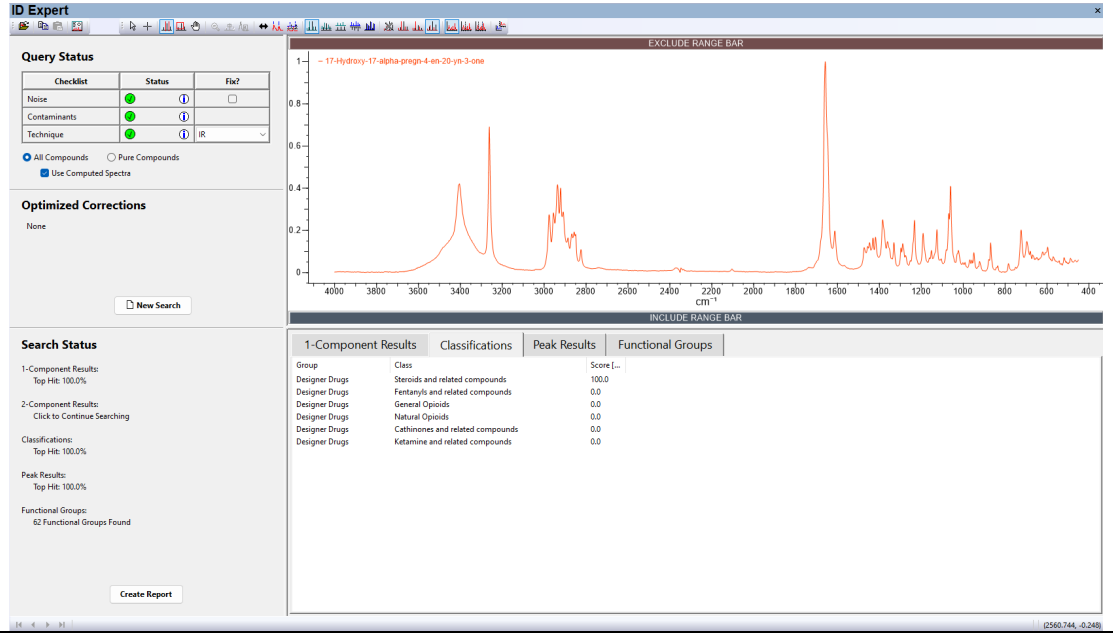
Classification models are a type of supervised learning used to label or categorize data into predefined classes or groups. These models predict the said label based on feature data. The model must be trained on labeled data or have an architecture that leads to the data being trained on to be self labelled by the model. Our models are binary classifiers that involve either a true or false result. The model also uses the neural network algorithm architecture, which has gained much popularity in recent years. The models have been optimized for FT-IR, Raman and GC-MS techniques within KnowItAll's ID Expert and MineIt applications.

ID Expert

Example File: 17-Hydroxy-17-alpha-pregn-4-en-20-yn-3-one.irf

	Action	Result
1	Open **ID Expert** (). If the **Open** window does not appear, then click **New Search** button () to manually launch the dialog window. 	The **Open** window appears for file selection: 

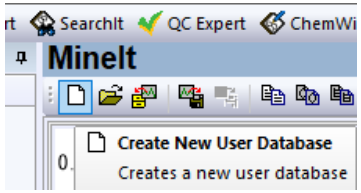
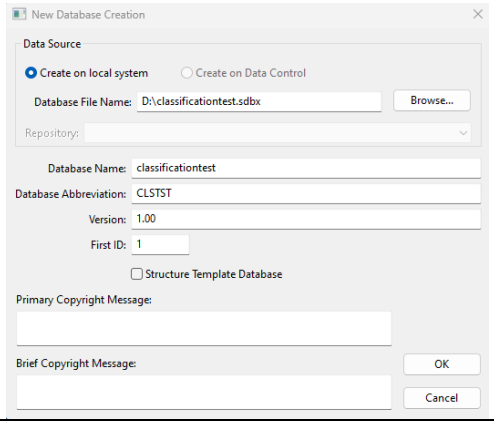
	Action	Result
2	Open **17-Hydroxy-17-alpha-pregn-4-en-20-yn-3-one.irf** in the **"C:\Users\Public\Documents\Wiley\KnowItAll\Samples\Mixture Analysis\IR Examples\Components"** folder. Click **Open**.	Immediately upon opening the file, it begins to search for a match using the licensed databases available to the user: 
3	Click **File > Settings**. Under the **Classification** tab of the **Settings** window, check the **Show All Classifications** checkbox. Click **OK** to save the changes. 	Upon clicking **OK**, the Settings window is closed.

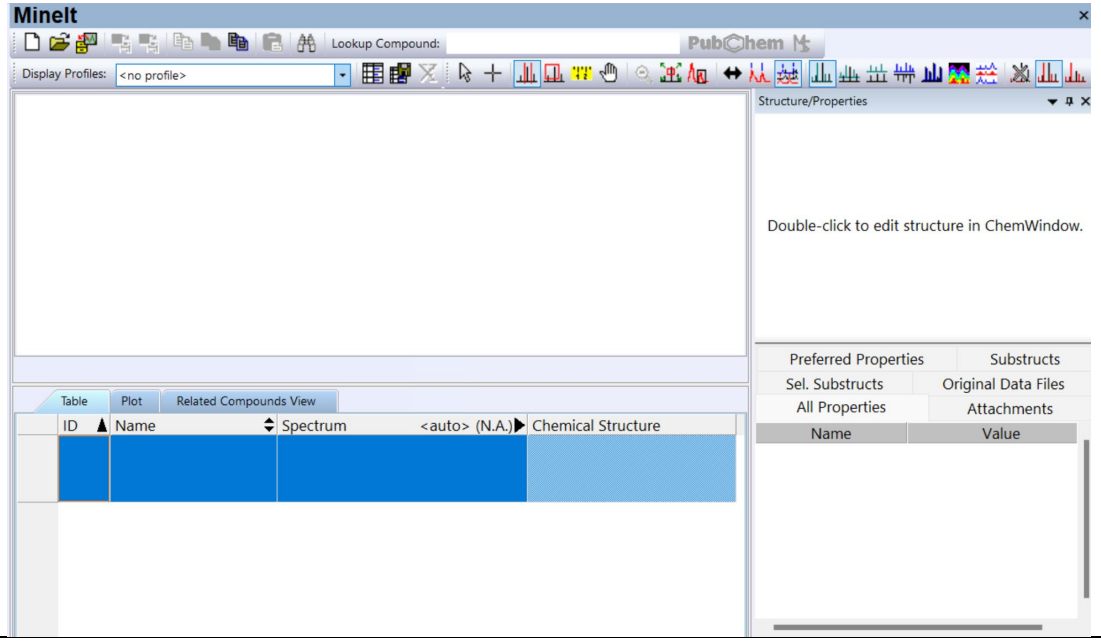
	Action	Result
4	Click the **Classifications** tab just right of the **Component Results** tab. *Note:* If there is a situation where there are multiple components detected in the sample (*i.e.*, >1), the tab is located to the right of the last component result as seen in the image below. 	**ID Expert** displays the classification results for the classification models available in the user license. In this case, the sample is identified to be a **Steroid**: 

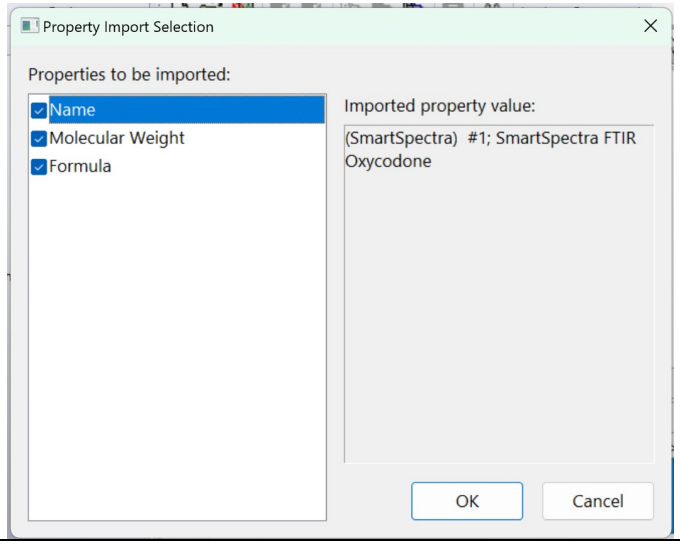
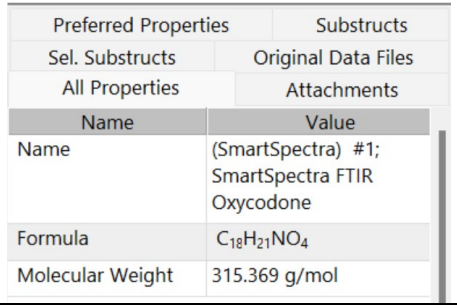
Classification Batch Property Calculation in MineIt

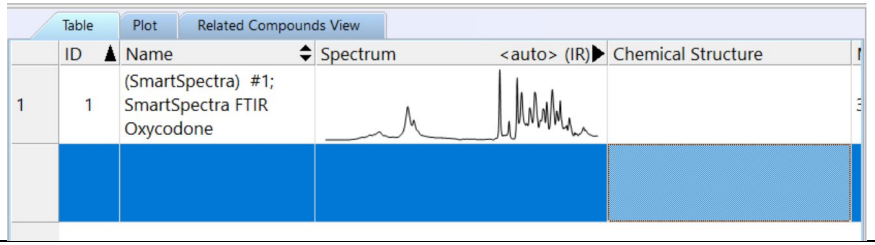
Example Files: SmartSpectraFTIROxycodone.irf, SmartSpectraRamanAndrosteroneacetate.irf

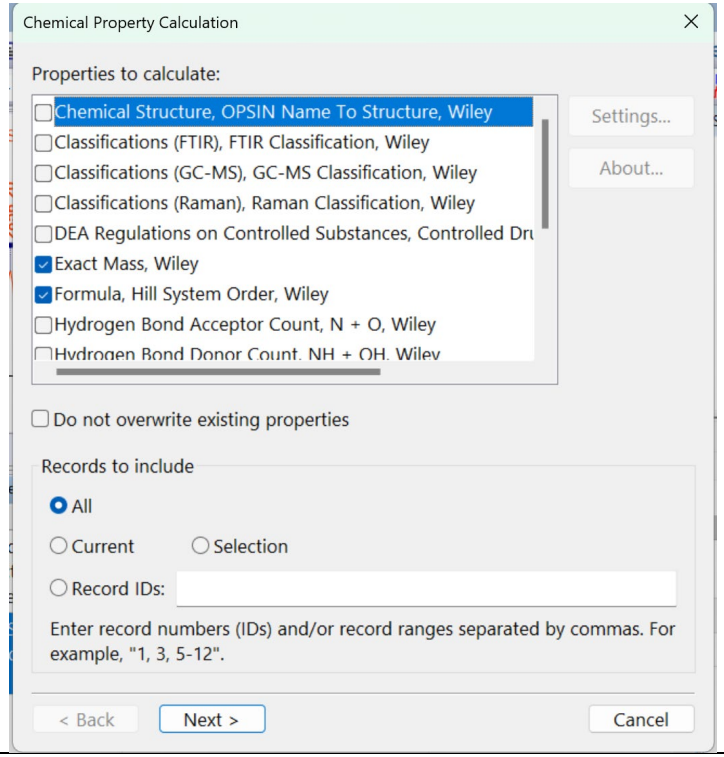
Batch property calculation is a calculation that will be applied to a series of compounds at one time. In this exercise, an unlocked user database will be created that to simulate executing a batch property calculation. This workflow can only be performed using unlocked (user) databases.

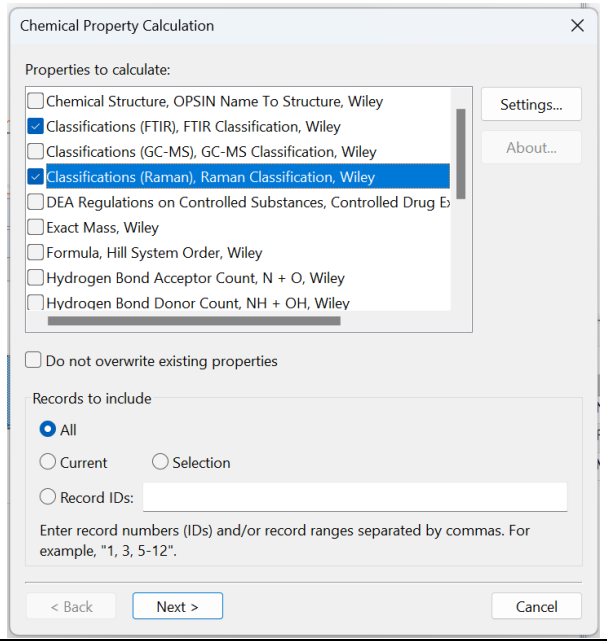
	Action	Result
1	Navigate to the **MineIt** application ( ). To use the **Batch Property Calculation** tools in **MineIt**, a user database must be created. *Note:* Creating a user database creates an unlocked database. Licensed KnowItAll databases are locked databases. Select the **Create New User Database** icon on the **Standard Toolbar.**	
2	In the **New Database Creation** window, choose a location to save the file by using the **Browse** button. Click **OK** to save. In the **New Database Creation** window, fill out the **Database Name** and **Database Abbreviation**.	The **New Database Creation** window is launched. 

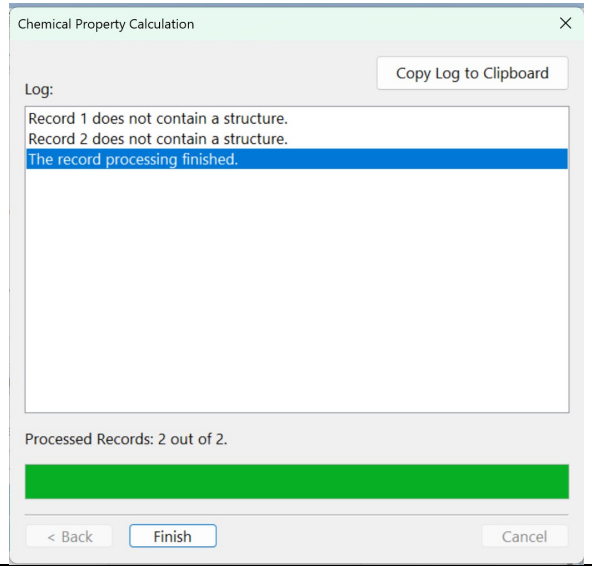
	Action	Result
3	Click **OK** to create the database.	A blank user database is opened in **MineIt**. 
4	Navigate to the **File** > **Import** to import spectral files.	The **Open** window is launched.

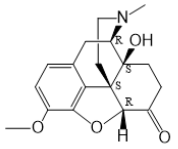
	Action	Result
5	Open **SmartSpectraFTIROxycodone.irf** in the **"C:\Users\Public\Documents\Wiley\KnowItAll\Samples\Computed Spectra\IR"** folder. Click **Open**.	The **Property Import Selection** window is launched: 
6	Click **OK** on the **Property Import Selection** window.	The **Property Import Selection** window is closed and the properties are imported in the **Structure/Properties** panel. 

	Action	Result
7	Click on the row below the first record in the **Table**.	The second row is shaded in the **Table**: 
8	Navigate to the **File** tab and select **Import** to import a second file into the user database.	The **Open** dialog window is launched.
9	Repeat steps 4 through 8 using **SmartSpectraRamanAndrosteroneacetate .irf** in the **"C:\Users\Public\Documents\Wiley\KnowItAll\Samples\ Computed Spectra\Raman"** folder.	The **Property Import Selection** window is launched. The Raman sample file is imported into the user database.

	Action	Result
10	Navigate to **Database > Batch Property Calculation**.	The **Batch Property Calculation** window is launched: 

	Action	Result
11	Click to select the **Classifications** models for IR and Raman: ▪ **Classifications (FTIR), FTIR Classification, Wiley** ▪ **Classifications (Raman),Raman Classification, Wiley"**. *Note:* It is a recommended to execute the models one at a time for their specific technique, but it is possible to run all classification engines at once on the different spectral records.	The **Chemical Property Calculation** window is shown below with the selected **Classifications** models selected: 
12	Under **Records to Include**, ensure **All** is selected. Then click **Next >** to execute the **Batch Property Calculations**.	

	Action	Result
13	Allow the **Batch Property Calculation** to execute entirely, where progress can be tracked using the green colored **Status Bar**.	Upon completion of the **Batch Property Calculations**, the **Log** box will read “The record processing finished”. The **Status Bar** will be entirely colored in green: 
14	Click **Finish** on the **Chemical Property Calculation** window to return to **MineIt**.	The **Chemical Property Calculation** window is closed.

	Action	Result
15	The resulting classification can be found in the **Structure/Properties** panel for the selected record.	Record ID 1 has results **Classifications (FTIR)** “General Opioids (100.0%)” and “Natural Opioids (100.0%)”. Preferred Properties | Substructs Sel. Substructs | Original Data Files All Properties | Attachments **Name** / **Value** Name: (SmartSpectra) #1; SmartSpectra FTIR Oxycodone Classifications: General Opioids (100.0%) Natural Opioids (100.0%) Formula: $C_{18}H_{21}NO_4$ Molecular Weight: 315.369 g/mol Record ID 2 has results **Classifications (Raman)** “Steorids and related compounds (100.0%)”. Preferred Properties | Substructs Sel. Substructs | Original Data Files All Properties | Attachments **Name** / **Value** Name: #2; SmartSpectra Raman Androsterone Acetate Classifications: Steroids and related compounds (100.0%) Formula: $C_{21}H_{32}O_3$ Molecular Weight: 332.484 g/mol
16	The final compound structures are provided in the **Results** cell for reference. These are the structural representations of the two spectra used in the batch property calculations added for comparison to show that they are classified correctly.	**Record ID 1:** (SmartSpectra) #1; SmartSpectra FTIR Oxycodone. 

	Action	Result
	To view all the different spectral techniques in the database spectrum pane, select the black triangle to the right of '<auto> (IR)' or '<auto> (Raman)' part of the pane. When the dropdown appears, select the technique you would like to view.	**Record ID 2:** #2; SmartSpectra Raman Androsterone Acetate.