WILEY

Validation Study: Wiley SmartSpectra IR Database

Abstract

In this study, we compare the performance of the *Wiley SmartSpectra IR Database* with the Wiley *Sadtler* libraries of IR spectra using the KnowItAll Analytical Edition 2024 software. This research was conducted to validate the accuracy and performance of the SmartSpectra library of computed spectra vs. empirical measured spectra.

To accomplish this, a validation test was developed and performed using the framework of an external validation study by evaluating the final computed spectra in a real-world library search scenario.

- A test of two hit list analyses of empirical replicate spectra was used to search against a library of empirical targets and the same set of replicate spectra was used to search against a library of computed spectra.
- These tests resulted in the correct corresponding match being in the top ten hits over 85% of the time for the library of computed spectra and over 94% of the time for the library of empirically sourced spectra.

These results demonstrate that the computed data performs the same test at a level comparable to empirical data searching for replicate matches.

Introduction

The *Wiley SmartSpectra IR Database*[1] is designed to provide those analyzing infrared spectra with greater coverage depth within a prescribed chemical space in which to identify general unknowns when an adequate empirical spectral library match is not available. The computed IR library, when used in conjunction with empirical libraries, can be used as a tool to determine the composition of an unknown spectrum and provide insight into the compound's associated structural makeup and constituent functional groups.

The goal of this validation study is to observe whether the computed library performs at a level comparable to traditional databases of empirically measured spectra. Owing to the novelty and experimental nature of the predicted library, an external validation study was completed[2]. This study was designed to evaluate the predicted library in a realistic scenario. The analysis of spectral search hit lists was used as a representative field test for how users

would experience using the library, as a test beyond the computational evaluation of a test set.

The use of spectral search software along with spectral reference databases is the backbone of characterizing and identifying unknown spectra. The results of spectral match comparisons or "hits" to reference databases are assigned a hit quality index (HQI) score to rank-order database matches based on how closely a spectrum matches the database's spectrum. In this study, we conducted a series of spectral searches using KnowItAll's[3] SearchIt application to compare spectra against both the empirical and predicted reference data sets.

Methods

Data

Using empirical "measured" reference data was determined to be the best way to create a test for the *Wiley Database of Predicted IR Spectra*[1], in addition to the standard train/test split used in computer modelling. The average hit list value was determined when searching for the replicate of empirical data within the vast catalog of Wiley's IR Spectral Database Collection[4]. To accurately compare with experimental data, a baseline was established for the comparison of replicate searching to determine how our experimental data performs in a search scenario (i.e., given a replicate, how good is recall from a library of empirical spectra versus recall from a computed library). This was completed with KnowItAll[3] software in an automatic batch method that exports a *.csv file of the results.

- The data for the replicates was sourced from the test set records, excluding any records which had the exact structure as records in the test set.
- This process then excludes any data that could have been used for the model to learn from as the model would have already known the outcome.
- The exclusion presents the model with unseen data for the most challenging test possible, similar to how this would be used in the field as a spectral library to be searched on to find unknowns.
- The records in the test set also could not contain replicates as there can only be one prediction per structure. This meant that there would still be replicates available in Wiley's data catalog for use in the replicate analysis. This is where the replicates were sourced from for both validation tests.

- The original test used 33 replicates from the model test set of 3,712 and the second test used 427 replicates from the model test set of 7,424 when the test set was expanded to 10% instead of 5% of the data used.

SearchIt

"SearchIt" is a KnowItAll[3] application used to search a spectrum, peaks, chemical structure, or property value against selected reference databases. For this test, we used spectrum searching alone. We used KnowItAll's correlation search algorithm, with and without employing KnowItAll's patented optimized corrections. Optimized corrections, among other functions, remove impurities from the spectrum, which can affect and often improve search matching.

Hit List Analysis

The hit list output was performed using a custom development tool. Two sets of different replicates derived from the test set were used in the analysis. These derived sets are designed to find the other's record through an exact structure search within the hit list and have a structure match in the predicted data set as well. For the validation test to function correctly, there can only be a single match of the compounds in each database to give accurate hit list results. In effect there are three databases:

1. one predicted test set,
2. the experimental test set, and
3. the target replicate set to search on the other two sets.

With the automated analysis, the spectra in both the predicted and empirical test sets were searched against the replicate target database with the alternative replicates inside. Initially, a spectral search was used to generate the hit list and then an exact structure search was performed on the hit list to find the position of the target replicate within the frame of each hit list. This was done to generate a baseline for predicted data against empirical data and empirical data against empirical replicates.

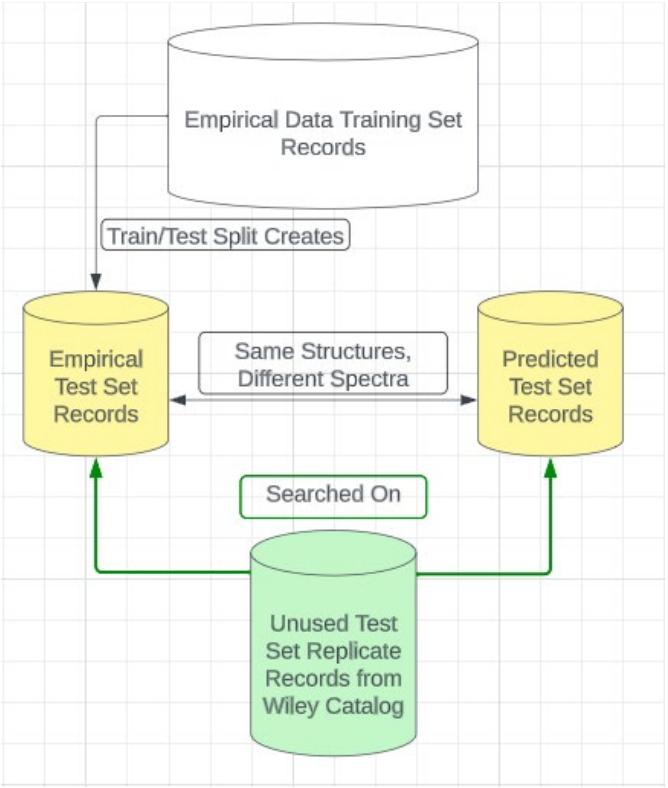

Figure 1: Flowchart of the training/test data split along with the visualization of the validation test

Results

First Validation Test

The results of the automatic validation show a high statistical similarity of hit list results between the empirical reference database and the predicted database.

The empirical replicates searched on the empirical test set database resulted in the correct reference spectrum within the top 10 hits 94% of the time, while also being the top hit 88% of the time. The empirical replicates searched on the predicted database resulted in the correct corresponding predicted spectrum in the top 10 hits 97% of the time and as the top hit 73% of the time. The results show that while empirical data is still superior to predicted data, predicted data can be an accurate tool to classify unknowns and increase the granularity of chemical space coverage.

Second Validation Test

The second validation test was run using the same parameters as for the first validation test but with slightly larger datasets to ensure the first test was accurate. The model test

set was expanded from 5% to 10% to allow for more replicates to be used in the analysis. The empirical replicates searched on the empirical test set database resulted in the correct reference spectrum in the top 10 hits 100% of the time and as the top hit 98% of the time. The empirical replicates searched on the predicted data set were in the top 10 hits 85% of the time and the top hit 58% of the time. The second validation test echoes the results of the original validation test as the empirical data performed incredibly well compared to the original results. The predicted data performed below the empirical data, though still at a high level.

When investigating the results of the second test, there were many spectral hits that had good HQI values greater than 75 but were not in the top 10 hits due to the structure type. For example, predicted long-chain hydrocarbons could have been overwhelmed in the hit list by many similar structured hydrocarbon spectra in IR spectroscopy. The spectral similarity for long-chain hydrocarbons is high enough that the slight differences in replicates could alter the expected hit list order quite easily. Another possibility for the lower hit list position could be the selection of replicates, where the first test had extremely varied replicates and the second test had extremely similar replicates.

The drop in performance may also be related to the amount of data used. When creating the second test set, more of the model's training data was removed in order to create a larger test set to find replicates. The second test set used 427 replicates from the model test set of 7,424 when the test set size was expanded to 10% of the data, instead of the typical 5% of the data. This increased the number of replicates significantly as there were only 33 in the 5% test set. Removing 5% of the training data to create a larger test set with more replicates which seems to cost in terms of performance. Luckily, the model used in prediction was the training set that used 95% of the data.

The validation hit list test was also run on the same datasets without using optimized corrections to see if the optimized corrections were strongly affecting the results for either dataset. The resulting hit list position was relatively the same as the testing with the optimized corrections toggled on, with minimal results affected. The replicates searched on the empirical data dropped to 91% for the correct result being in the top 10 hits and similarly scored 88% as the first hit. The record was in the top 10 hits 94% of the time and the top hit 64% of the time. The takeaway here is that with optimized corrections off, the empirical data for all purposes, tested the same, while the predicted data enjoyed better results with the corrections on. However, this only affected the top hit results, while the top ten hit results remained relatively the same.

Results Table:

	Average Hit List Position	Top 10 Hit Percentage
Validation Test 1: Replicates Searched on Empirical Test Set	3.97	94%
Validation Test 1: Replicates Searched on Predicted Test Set	2.94	97%
Validation Test 2: Replicates Searched on Empirical Test Set	1.11	100%
Validation Test 2: Replicates Searched on Predicted Test Set	9.45	85%

Conclusion

The *Wiley SmartSpectra IR Database* required a true test of how the library would perform in the field. The validation study was designed to be an external test of the finished product, a test beyond the computational evaluation of a test set. The ideal test was an analysis of the hit list generated from searching a spectrum against a spectral library for an exact structure match. The test set data was then used to create the external automatic validation tests which showed that predicted results can work nearly as well as empirical results. The predicted database searched comparably to the empirical data, with both having the correct match in the top ten hit list results around 95% of the time in the first test and both being above 85% in the second test. While it is apparent that empirical data is still superior, we can see that the predicted data performs at a level similar to empirical data.

These findings validate the utility of the predicted data sets within the *Wiley SmartSpectra IR Database* as a valuable complement to empirical reference datasets for broadening the accessible chemical landscape. Predictive libraries, particularly when constructed from extensive and high-quality empirical reference datasets such as Wiley's extensive spectral collections (including premium Sadtler spectra), demonstrate performance levels closely approaching that of empirical datasets. Consequently, they can be effectively integrated into the analytical workflow.

References

1. *Wiley SmartSpectra IR Database*. John Wiley and Sons, Inc. 2024 https://sciencesolutions.wiley.com/machine-learning-and-prediction-at-wiley-science-solutions/(last accessed 2024-07-23).
2. Ho, S.Y., Phua, K., Wong, L., & Goh, W.B.G. Extensions of the external validation for checking learned model interpretability and generalizability. *Patterns* **2020**, 1(8). DOI: https://doi.org/10.1016/j.patter.2020.100129 (last accessed 2024-07-23).
3. KnowItAll Analytical Edition 2024, Release 24.0.59.0. https://sciencesolutions.wiley.com/knowitall-analytical-edition-software/, John Wiley and Sons, Inc. 2024 (last accessed 2024-07-23).
4. KnowItAll IR Spectral Database Collection. John Wiley and Sons, Inc. 2024 https://sciencesolutions.wiley.com/solutions/technique/ir/knowitall-ir-collection/ (last accessed 2024-07-23).