WILEY

Validation Study: Wiley SmartSpectra Vapor Phase IR Database

Abstract

The Wiley SmartSpectra Vapor Phase IR Database[1] is the latest computed spectral library release from Wiley Science Solutions. This library uses the same model architecture as the Wiley SmartSpectra IR Database[2] but has been optimized for Vapor Phase IR (VPIR) specifically, as it was trained only with this specific subset of IR data. To accurately evaluate the predictions, a validation test was created and conducted using the framework of an external validation study.

The aim of the study was to test whether the computed data could perform at the same levels as replicate data, but there was not enough replicate data to perform this style of evaluation. Instead, a secondary test was conducted using a set of data from Sigma-Aldrich to which model had not previously been exposed to. This increased our model test set size of 399 records to 4,542 records including the Sigma-Aldrich test set (included in the KnowItAll IR Spectral Database Collection[3]).

This test was then repeated using the same data as well as all of Wiley's Vapor Phase data[3], increasing the number of records being searched on to 15,731 unique records. The results showed that the prediction model was in the top ten hits 88% of the time across all three validation tests. This external validation study[4] showed that the predicted data performs extremely well when used for searching across various testing parameters and when searched on large amounts of data.

Introduction

The Wiley SmartSpectra Vapor Phase IR Database was published to provide additional compound coverage depth within the existing chemical space of Wiley's licensed VPIR data. This additional coverage is designed to identify unknowns when an adequate empirical spectral library match is not available. Therefore, in conjunction with empirical VPIR empirical spectral libraries, this predicted spectral library can be used effectively as a tool for unknown VPIR spectral classification and characterization. To ensure that this database was ready for customer use, a validation study was conducted to observe if the predicted VPIR data performed similarly to empirical VPIR replicate data.

Since the VPIR library used the same model as the predicted IR library, the initial idea was to try and develop a replicate analysis similar to that which was developed for the previous predicted library. Unfortunately, there were not enough replicates to perform this style of test at a quantifiable level. The replicate testing revealed only 37 compounds with replicates

that were not in the training set and available to be used in this analysis. To mitigate the lack of replicates, the size of a secondary test set was increased using data the model had previously not been exposed to. The same test with the larger test set was also run against all our empirical libraries to simulate searching for an unknown using the VPIR predicted library. This study was designed for testing the predicted spectra in a realistic scenario. The analysis of the spectral search hit lists was used as a representative field test to mimic how users would experience the library as a finished product, a test beyond the model's own test set.

Spectral search software in combination with empirical spectral databases are used as the basis for characterizing and identifying unknown spectra. The spectral matching result comparisons or 'hits' to reference databases are assigned a hit quality index (HQI) score to rank the database matches based on the proximity of the searched spectrum to the database's spectral match. For this external validation test, KnowItAll's spectral searching application SearchIt was used to compare spectra against both the empirical and predicted datasets.

Methods

Data

The best way to create a test set for the VPIR prediction model was determined to be the inclusion of the Sigma-Aldrich Library of Vapor Phase FT-IR Spectra, as these records were not used in the creation of the model. This Sigma-Aldrich Vapor Phase IR library was then evaluated for records that might be in the VPIR training set and these records were then removed. The average hit list value was determined when searching for the replicate of empirical data within the catalog of Wiley's VPIR spectral database collection. To accurately compare with experimental data, a baseline was established for the comparison of replicate searching to determine how Wiley's experimental data performs in a search scenario (i.e., given a replicate, how good is the recall from a library of empirical spectra versus recall from a predicted library). This was accomplished using KnowItAll[5] software in an automatic batch method that automatically exports a csv of the results. The replicate data was sourced using the model's test set records, the matched chemical structures in the test set were predicted to be used in a realistic spectral search simulation. The records in the test set cannot have structures that could be in the training set, as the compounds, even if they had a replicate, cannot be in both the training and test set. This infers that there could have been many replicates in Wiley's data catalog for replicate analysis use. The replicate testing using hit list analysis was run with a sourced combination of Sigma Aldrich and Wiley data, resulting in 37 compound pairs from the model test set of 399 records. Since this amount of data was so low, the use of the Sigma-Aldrich VPIR library without compounds that appear in the training set and that are within the model boundaries were extracted for a hit list analysis test. This library was then analyzed for the compounds that were within the model's chemical space boundaries and those records not in the overlap were removed.

This resulted in 4,542 records that were within the chemical space from the Sigma-Aldrich VPIR library. The 4,542 structures were then predicted to create a new database to be searched on by the original Sigma-Aldrich VPIR library's spectra. The tertiary test used both the Sigma-Aldrich empirical and predicted libraries in another hit list analysis test with Wiley's additional 11,189 records added to the databases being searched on to give a larger data pool, simulating a realistic spectral search use case.

SearchIt

"SearchIt" is a KnowItAll 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 KnowItAll 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 matching structure 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.

To supplement the low amount of replicate data, additional test sets were sourced to run a hit list analysis on. The analysis hinges on the same principles as the replicate analysis, but with only 2 data sets. The first data set that does the searching is the original empirical spectrum and the second data set is the computed SmartSpectra dataset that is searched on. This can be modified to increase the difficulty of the analysis by adding more empirical data to increase the chances that the corresponding records do not find each other. This test would apply the same principles of having a corresponding structure match to find the hit list position while seeing if the spectra were similar to each other based on HQI.

With the test set automated hit list analysis, the spectra in the predicted test sets were searched on by the empirical database containing the original records. 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's structure within the frame of each hit list, as previously mentioned. This was done to generate a baseline for the predicted data against empirical data and the empirical data against empirical replicates.

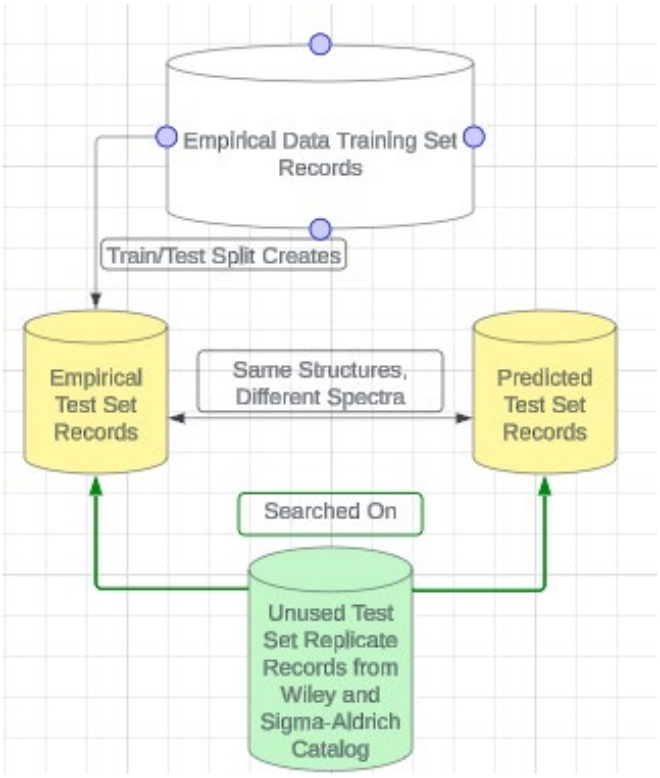

Figure 1: Flowchart of the training/test data split along with the visualization of the validation test

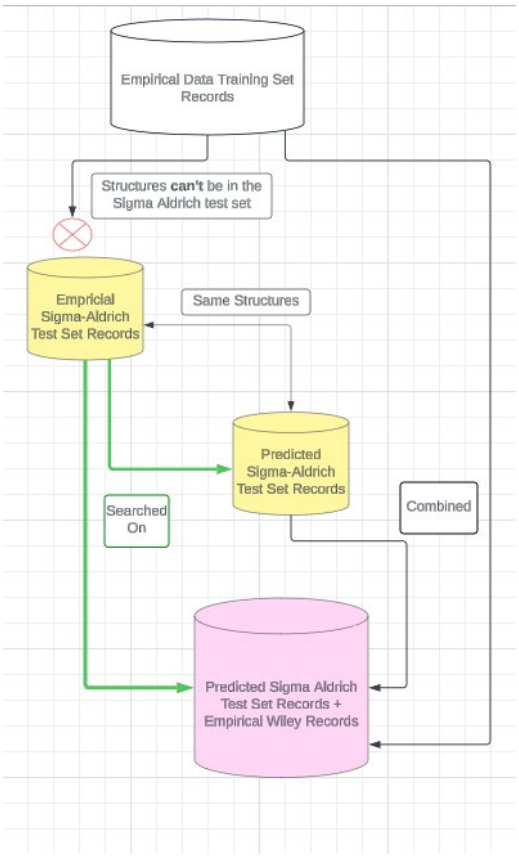

Figure 2: Flowchart of the Sigma-Aldrich data set validation test

Results

Replicate Analysis Validation Test

The replicate analysis test resulted in all but one 37 corresponding spectra finding each other, most likely due to the low number of records. Therefore, the two tests of replicates searched on the original empirical records and replicates searched on the predicted records resulted in the same statistics for first hit rate and top ten hit rate. The only statistic that was relatively different was the HQI for the hit's spectral matching, of which the replicates

searched on the empirical data resulted in an average HQI of 94.66, while the replicates searched on the predicted data resulted in an average HQI of 84.91. This shows that the model is quite close to the quality of replicate data, but also not quite as good as empirical data, replicate or not.

The Model's Validation Test Set

The original test set was used to evaluate the model with data that was not in the training set. The empirical test set was searched on the predicted test set to evaluate the prediction engine's spectra. The test set contained 399 records and the hit list testing resulted in an average hit position of 4.01, while also producing a top hit percentage of 83% and a top ten hit percentage of 93%. The average HQI for the corresponding spectral match was 80.79 for the model test set. The results from this hit list testing were promising, but since there was not a lot of data, further test sets were created to further evaluate the model's performance.

Sigma-Aldrich Validation Test Set

To supplement the original test set from model generation, the Sigma-Aldrich Library of Vapor Phase FT-IR Spectra[3] was used to acquire compounds that were not in the training set for hit list testing. Compounds outside of the model's chemical were removed, as well as any replicates, for the hit list analysis to work. The final count of Sigma-Aldrich data was 4,542. These records were then used to create another database where the model predicts the spectra for all 4,542 records. The empirical Sigma-Aldrich data was then searched on the predicted database to analyze the model's predicted spectra using hit list analysis. This test resulted in an average hit position of 7.6, with a first hit rate of 70% and a top ten hit rate of 90%. The average HQI for the Sigma test set was 86.08, higher than the model's own test set. The hit rate results were very similar to that of the model test set, only dropping in the first hit percentage, yet maintaining a close percentage in the top ten hit rate. The results of this test were promising as well, and only declined marginally with the addition of over 10x the data as the first test set.

Sigma-Aldrich Validation Test Set with Wiley Empirical Data

To ensure that the predictions on the Sigma-Aldrich test set would perform similarly with a larger pool of data, more data was added to the databases to be searched on and the test was rerun. There were 11,189 compounds added to the databases being searched on from Wiley's Vapor Phase IR catalog. These records were intended to create more of a challenge for the hit list analysis test. By introducing more spectra to the spectral search, similar looking spectra could change the results if the compound only differs by a single atom or bond. This version of the test was the hardest for the predicted Sigma-Aldrich test set, as the amount of data increased by 246%. The 4,542 records of the empirical Sigma-Aldrich test set were then searched on the 15,731 records which included the predicted Sigma-

Aldrich test set. The test resulted in an average hit position of 8.94 and a first hit rate of 69% and a top ten hit rate of 88%. Even with the additional data added to the databases being searched on, the Sigma-Aldrich predicted test set resulted in extremely comparable numbers. The hit rates only dropped slightly from the previous test and the hit list position on average only increased 1.34 positions.

Results Table:

	Average Hit List Position	Top 10 Hit Percentage
Validation Test 1: Replicates Searched on Empirical Test Set	1.05	100%
Validation Test 1: Replicates Searched on Predicted Data	1.05	100%
Validation Test 2: Wiley's VPIR Model Test Set	4.01	93.5%
Validation Test 3: Sigma-Aldrich VPIR Library Test Set	7.60	90.3%
Validation Test 4: Sigma-Aldrich VPIR Library Test Set with additional Wiley Data included in the searched databases	8.94	88%

Conclusion

These validation tests were designed to serve as an external validation test, to see how the predictions would behave in a field-like scenario and to test the model's prediction capabilities against compounds that the model had not seen before. The two test types of testing were conducted, with the validation set hit list testing repeated three times with differing parameters. The first test was replicate analysis, observing whether a replicate of a compound would perform at the same level as the prediction of the compound. This was tested using hit list analysis generated through KnowItAll's SearchIt, matching spectra based on HQI and relaying the position of the exact structure match in the hit list. With the limited amount of unseen replicate data used, the predicted data scored the exact same scores as the replicate spectra, fully passing the test.

However, since this was a small amount of data, it was decided to run multiple test sets with increasing size of matched spectra and datasets searched on. These tests had excellent results as well. All 3 test sets had a top 10 hit percentage above 88%, even when the pairing size increased 10x and when the data pool to be searched on increased by 246%. These tests show that the predicted data performs at a level similar to empirical data, though it is apparent that empirical data is still superior.

The performance level of the model's predictions validates the utility of the predicted data sets within the Wiley SmartSpectra Vapor Phase IR Database collection as a complementary library to the empirical reference datasets for broadening the searchable chemical space. The predicted libraries, when created from high quality empirical reference spectral databases such as Wiley's Vapor Phase IR spectral collection, demonstrate a high level of performance approaching that of empirical databases. These libraries have shown the ability to characterize and classify more unknowns by enhancing the coverage within the bounds of Wiley's Vapor Phase IR chemical space.

References

1. Wiley SmartSpectra Vapor Phase IR Database. John Wiley and Sons, Inc. 2024 https://sciencesolutions.wiley.com/machine-learning-and-prediction-at-wiley-science-solutions/ accessed 2024-07-18)
2. Wiley SmartSpectra IR Database. John Wiley and Sons, Inc. 2024 https://sciencesolutions.wiley.com/solutions/technique/ir/wiley-database-of-predicted-ir-spectra/ (accessed 2024-07-18)
3. KnowItAll IR Spectral Database Collection. John Wiley and Sons, Inc. 2024 https://sciencesolutions.wiley.com/solutions/technique/ir/knowitall-ir-collection/ (accessed 2024-07-18)
4. 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
5. KnowItAll Analytical Edition 2024, Release 24.0.59.0. https://sciencesolutions.wiley.com/knowitall-analytical-edition-software/, John Wiley and Sons, Inc. 2024 (accessed 2024-07-18)