WILEY

Validation Study: Wiley SmartSpectra Raman Database

Raman Spectral Prediction Using Neural Networks

Abstract

As the demand for Raman spectra increases, Wiley Science Solutions has embarked on the challenge to develop a computed library of Raman spectra within the constraints of our current empirical Raman collection chemical space. The Wiley Raman SmartSpectra Database is a computed spectral library that offers access to 33,163 computed records.* This database is designed to expand the chemical space covered as a supplement to our comprehensive collection of empirical, measured Raman spectra[1]. It can help users further elucidate the possible identity of unknowns when an empirical library match is not available.

The model uses a custom-built chemical structure fingerprint as the input and predicts a Raman spectrum as the output. This prediction model was evaluated by an external validation test using JASCO data[2], to accompany the model's test set, in which the JASCO empirical data were searched against the predicted version of these records. The results returned the corresponding target structure in the top ten hits over 82% of the time, as compared to the model's own test set of 91%.

** Represents the number of spectra in the collection as of the publication of this study.*

Introduction

The recent increase in demand for Raman spectra has fueled the development for more experimental Raman analysis of chemical compounds. Wiley already publishes one of the largest Raman collections available, the KnowItAll Raman Spectral Database Collection, accessible through Wiley's software, KnowItAll 24.0.59.0[3]. With the goal to expand Wiley's Raman coverage within the bounds of the existing Raman collection, a predictive model was created to increase the number of characterized compounds. The model was able to improve the coverage of Raman within the current collection's chemical space, filling in some of the theoretical missing gaps with computed spectra. To confirm and increase the accuracy of the Raman predictions, validation tools were developed to carry out additional tests on the predictions.

The computed Raman library was designed to help identify an unknown when an empirical library match of sufficient quality is not available to the user, i.e., in the theoretical case of a true unknown. When used in tandem with the empirical Raman libraries, the computed library can be used as an aide to help characterize or classify the structural composition of an unknown spectrum by providing information about the structural makeup, such as functional groups and chemical structure backbone. However, the computed Raman library is not a substitute for empirical Raman libraries. Although there are cases where the computed Raman library will fully characterize a compound with a computed Raman spectrum, typically results will only provide part of the identity of the unknown compound.

The computed library should be used, after searching the empirical Raman libraries, to gather additional information about the unknown compound or if the empirical search results were inconclusive. The computed library's purpose is to characterize compounds not previously available in our catalog, with the compound still falling within the bounds of our chemical space. This provides the benefit of access to a more comprehensive selection of compounds and to improved overall coverage and selectivity. This paper presents the computational methods used to develop the model and library, while also exploring the quality of the predictions through automatic validation using HQI (hit quality index) and manual validation by subject matter experts.

Methods

Model Architecture

The model takes a custom-built chemical structure fingerprint[4] as the input and predicts a Raman spectrum as the output. The architecture consists of an input layer followed by multiple dense layers with ReLU (Rectified Linear Unit)[5] activation. Dropout regularization[6] is applied after each dense layer[7] to prevent overfitting[8]. The final layer uses Sigmoid activation[9] to produce the predicted spectrum. The model is constructed using the Keras library version 2.10.0[10] and TensorFlow version 2.10.1[11].

Hit List Automatic Validation

Due to the novelty and experimental nature of the computed library, an external validation study[12] was performed. The model naturally produced a traditional test set for initial analysis, which under typical circumstances would be sufficient. However, because this is a novel computed library, it was decided to employ an additional external validation test[13]. To evaluate the computed library in a realistic scenario, analysis hit lists were used as an accurate field test as to how users would experience using the library.

Data

Using experimental reference data was determined to be the best method to create a test for the computed library, in addition to the standard train/test split used in computer modelling. JASCO allowed the use of their Raman spectra database to test the ability of the prediction engine. The JASCO records that were not in Wiley’s library were used as a

prediction set for testing the search recall. The Wiley predicted version of the JASCO set was added to a catalog of Wiley Raman databases for testing. The searched databases contained around 8,000 records and the JASCO set contains 328 records within the model's chemical space. This set of tests was created with the original model's test set, the JASCO test set alone, and the JASCO test set combined with Wiley's Raman standards. This was accomplished using KnowItAll software in an automatic batch method that automatically exports a csv of the results.

SearchIt

"SearchIt" is a KnowItAll application used to search a spectrum, peaks, chemical structure, or property value against selected databases. For this test, we used spectrum searching alone. We used the KnowItAll correlation search algorithm[14], with and without employing KnowItAll's patented optimized corrections. Optimized corrections, among other functions, removes impurities from the spectrum, which can affect and often improve spectral 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 computed 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 computed 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 two 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 automated analysis, the spectra in both the predicted and experimental test sets were searched against the replicate target database containing the alternative replicates. 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 an external test for computed data using JASCO empirical data.

Spectrum/Structure Validation Model

Wiley also developed a spectrum/structure validation model for the purpose of validating spectrum/structure pairs in the reference data. This model was also used on computed spectra, which gave another accuracy metric to identify and remove any poorly computed data that were outside the bounds of the underlying training set. The validation scores will be included with the computed library as an additional accuracy metric for user confidence. The validation model performs at a high level, producing test set scores at or above 85% for accuracy (85%), precision (79%), recall (93%), and F1 (85%).

This validation model was trained on Wiley's Raman spectral data records and the associated chemical structure converted into a chemical fingerprint with each record. It was impossible to validate those without a structure as there is no comparative metric to evaluate the initial spectrum. This model returned a statistic (0.0-1.0) for the probability that the spectrum and structure fingerprint are correct for each other. The model was trained in different oversampling techniques[15] to create a resultant database for each outcome by type of oversampling method (e.g., spectral smoothing, added noise, etc.). These methods for adding replicate data were the best attempt at creating realistic scenarios for the model to learn of/from differing data that is technically correct. Here, the goal was to utilize examples of Raman spectra with slight contamination or with spectral measurements from differing instruments, as it is known that instruments of different brands can often provide different intensities[16]. These variations on the original data give our model varied datasets that allow for an accurate validation of the spectrum/structure pairing.

Oversampled false match datasets were created with a similar concept behind them, including changes to the spectrum/structure pairing for the model to evaluate both properties before giving a score on whether the pairing can possibly be related to one another. This includes creating false oversampling techniques by changing either the spectrum or structure fingerprint, or both. These true negative oversampling methods allow the model to learn with 'bad' data as well, which is essential to the model's accuracy.

Validation and Review: Experimental spectra searched on the predicted test database

Two test sets were derived from the 5% split test set that was generated during the model creation. Each set consisted of 537 records, where one set contained real experimental records and the other contained computed records. In the MineIt application, the first 100 spectral records from the empirical data test set were used in the evaluation. Each of the 100 records were transferred from the empirical data test set to the SearchIt application.

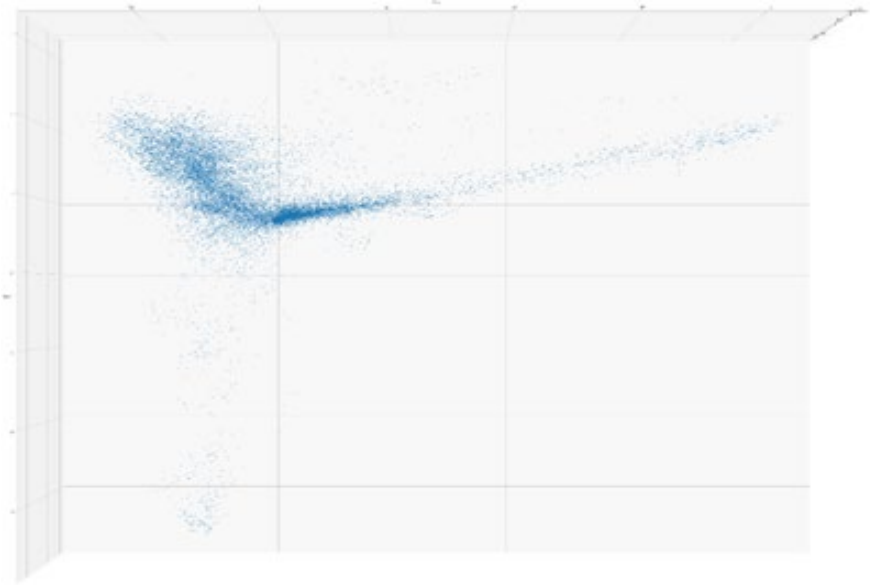
The predicted data test set was selected as the user database to be searched. The chemical structure was then removed, leaving only the spectrum for conducting the search. The following parameters were used: the search algorithm was set to "Correlation", "Optimized Corrections" was also selected as most people would use this in their search, and in the Advanced Settings, "Remove Duplicates" and "Remove Replicates" were deselected. The hit list size limit was set to 100. The spectral search was performed, and the hit list number was observed for the exact structure match from the predicted test set hit list. Other hits

on the hit list were also observed to note similarities and differences in the predicted results.

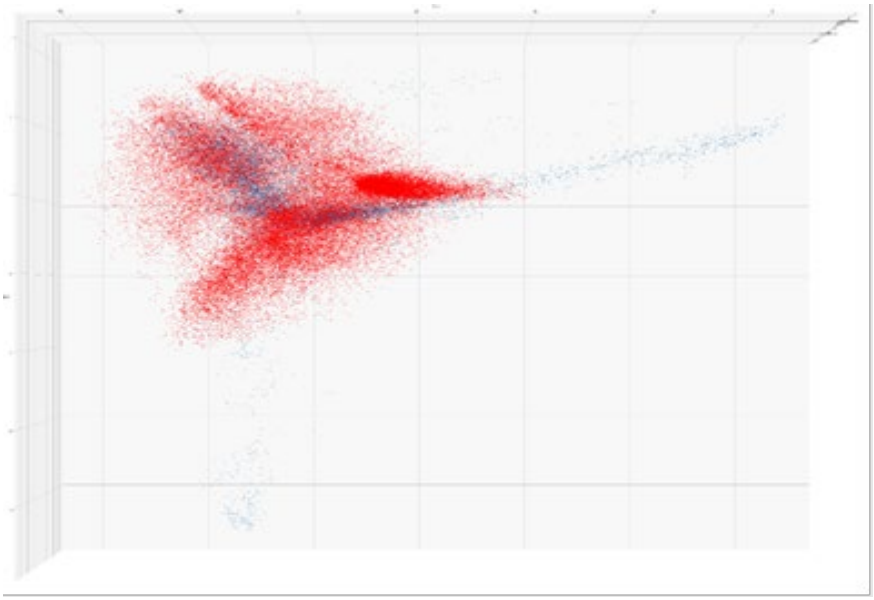
Results and Discussion

Chemical Space Coverage

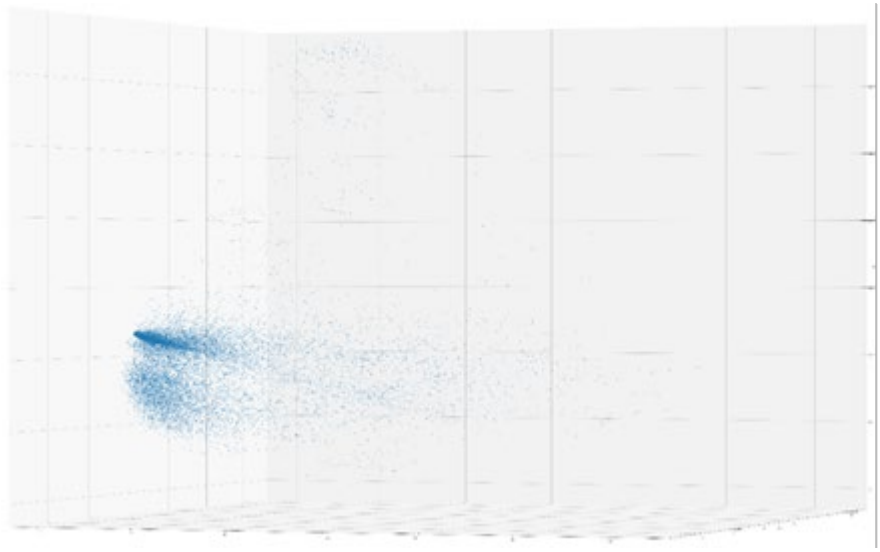
Elevation 180, Azim 0, empirical data only



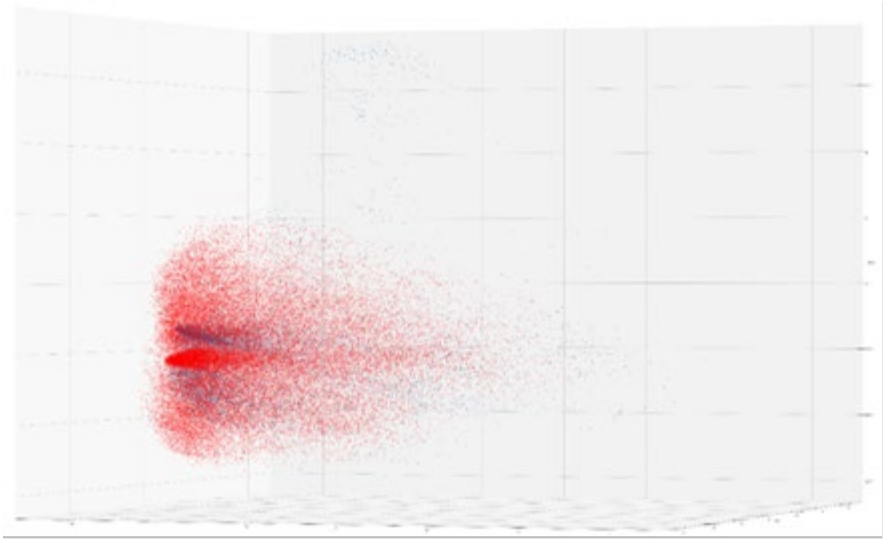
Elevation 180, Azim 0, predicted data added



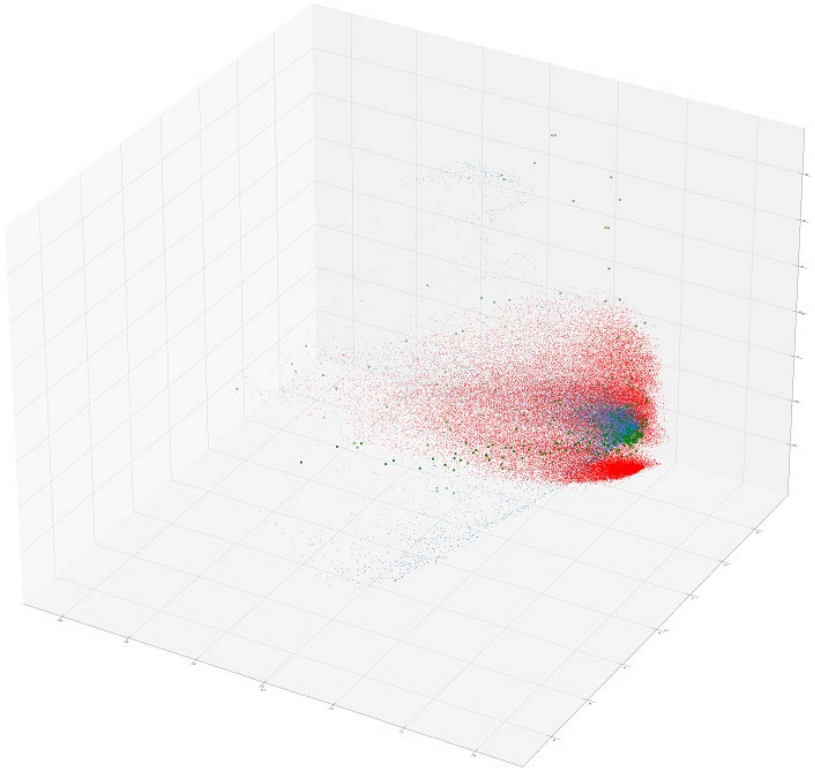
Elevation 0, Azim 290, empirical data only



Elevation 0, Azim 290, predicted data added



Elevation 30, Azim 120, predicted data and JASCO data (green + enlarged) added



Chemical Space Coverage

Overall

The hit list test searches empirical data against the computed database. The first test returned good results, in which the empirical records found their corresponding computed record in the top 10 hits over 77% of the time. The second test used the same parameters but involved JASCO data to confirm the original test, resulting in the corresponding predicted record being in the top 10 hits over 82% of the time. For the third test, the second test with JASCO data was duplicated but with an additional 8,000 records added from the Sadtler standards Raman library suite (part of the KnowItAll Raman Spectral Database Collection) to simulate a larger pool of data for searching. This test resulted in the same hit list statistics as the previous test using JASCO data, in which the corresponding predicted record was in the top 10 hits over 81% of the time when searched against 24x the amount of data.

Replicate Analysis

The replicates hit list analysis resulted in the predicted version of the replicates having an average position of 10. The replicates had a first position hit rate of 38% and a top 10 hit rate of 74%. This is due to the lack of replicates in our data. There were only 46 replicates in Wiley's Raman libraries, therefore if there were any discrepancies their value would be increased compared to a more populated set of replicates. Since there were so few replicates and many of the records were closely related structures, the results were underwhelming. Additional hit list analyses were done to mitigate this lack of replicate data. To test the model's capabilities further there were two separate test sets created to mitigate the lack of replicate data.

Test Set Analysis

The model's test set was used in this analysis. Based on the model's data split of 95% training and 5% test, there were a total of 537 records available for this validation test. The 537 records had their structures removed and then computed in a separate database for this purpose. This computed database was then searched on in an automated fashion using a custom build of KnowItAll. The resulting analysis showed that the predicted data was the top hit 64% of the time and in the top ten 91% of the time.

JASCO Hit List Analysis

JASCO allowed the use of their Raman database to test Wiley's Raman prediction model. The database was analyzed to select records not available in Wiley's Raman computed model's training library, thereby excluding a few hundred records. These records were then computed using the model and any records outside of the structure space were removed, leaving 328 records as a secondary test set for hit list analysis. After finalizing the JASCO test set, the predicted versions of the records were first run on their empirical counterparts. This initial hit list test resulted in a 58% rate of first hit and was in the top 10 hits 82% of the time.

A second test using the JASCO dataset was performed by adding more databases to simulate a realistic scenario in which a user may include more databases. For this test, there were more Wiley empirical databases added to the search list along with the SmartSpectra computed JASCO database for a total of 8,536 records. This simulated a large database for Raman, with the target records only representing 4.1% of the data. The result of this test was a 56% rate of a first hit and an 81% top ten hit rate for the corresponding spectrum. The average hit position was 15th but only 19% of the records were outside of the top 10, indicating that the model's average hit position was being negatively influenced by large hit positions when outside the top 10. When the hit results that were outside the hit list were removed, the percentages increased drastically. Discarding these extreme data points decreased the average hit position of the JASCO test set searched with Sadtler data to 5.7, increased the first hit percentage to 61% and increased the top ten hit percentage to 89%.

	Average Hit Position	Top Ten Hit %
Model Test Set	6.1	91%
JASCO Test Set	13.1	82%
JASCO Test Set with Sadtler Empirical Data	14.6	81%
JASCO Test Set with Sadtler Empirical Data (Outliers Removed)	5.7	89%

Manual Subject Matter Expert (SME) Validation

The top 10 spectra hits of the hit list from the predicted data test set were very good spectral matches for the empirical data test set spectra. The functional groups and fingerprint region were well generated, resulting in the high hit list position. Additionally, the remaining percentage of spectra that had a hit list position greater than 100 still contained some good spectral matches that had poor hit list position due to similar compound coagulation. Overall, the analysis was good as there were high top ten percentages (92% and 77%). Although the rate of first hit declines from the test set compared to the total empirical collection, there were high percentages for the 2nd and 3rd hit (14% and 9%), which infers the top 3 hit positions were returned 66% of the time.

Summary of Results	**Experimental Data Test Set Evaluation** **• Searched against only the Predicted Data Test Set Raman Database**
Hit List Record Number	**Top 10: 99 records** • Hit #1 = 73 • Hit #2 = 12 • Hit #3 = 4 • Hit #4 = 2 • Hit #5 = 3 • Hit #6 = 2 • Hit #7 = 1 • Hit #9 = 1
Hit List Record Number	**Between Hit List Records #11 and #100: 7 records** • #16, #16, #17, #20, #20, #28, #30
Hit List Record Number	**Not in Top 100: 2 records**

Summary of Results	**Experimental Data Test Set Evaluation** **• Searched against only the Predicted Data Test Set Raman Database and Experimental Raman Databases (~17k)**
Hit List Record Number	**Top 10: 83 records** • Hit #1 = 46 • Hit #2 = 15 • Hit #3 = 10 • Hit #4 = 4 • Hit #5 = 4 • Hit #6 = 2 • Hit #7 = 1 • Hit #8 = 1
Hit List Record Number	**Between Hit List Records #11 and #100: 13 records** • #19, #20, #21, #22, #26, #38, #39, #44, #45, #47, #63, #66, #78
Hit List Record Number	**Not in Top 100: 12 records**

High Molecular Weight Prediction Testing

A tertiary test was conducted to determine whether the higher molecular weight predictions are as good as the lower molecular weight predictions. This test was designed as over 80% of the data in Wiley's Raman collection are under 400 g/mol. An analysis was conducted to determine if a hard cutline of molecular weight should be implemented on the predicted library. There were 50 records above 384 g/mol from the Raman test set used for the analysis. The procedure for searching was done by transferring the empirical corresponding spectrum to SearchIt, removing the chemical structure, setting the algorithm to correlation, and limiting the hit list size to 100. Then, the search was performed first against the predicted test spectra only (537 records), followed by the predicted test spectra and Wiley's collection of Raman spectra (~17k records). The lowest weight was 384 g/mol and the highest was 695 g/mol. The test had great results against only the predicted test set, where there was an 80% first hit rate, while testing against the predicted test set and the empirical Raman collection returned a lower first hit rate at 42%. The top ten hit percentages were very similar to each other as the analysis against only the predicted test set had a result of 94%, and the analysis against the empirical and predicted test set had a result of 88%, which is very comparable.

Summary of Results	**<u>Experimental Data Test Set Evaluation</u>** **• Searched against only the Predicted Data Test Set Raman Database (537 records)**
Hit List Record Number	**<u>Top 10:</u> 47 records** • Hit #1 = 40 • Hit #2 = 2 • Hit #3 = 3 • Hit #4 = 1 • Hit #6 = 1
Hit List Record Number	**<u>Between Hit List Records #11 and #100:</u> 2 records** • #16, #61
Hit List Record Number	**Not in Top 100: 1 record**

Summary of Results	**<u>Experimental Data Test Set Evaluation</u>** **• Searched against the Predicted Data Test Set Raman Database and Experimental Raman Databases (~17k records)**
Hit List Record Number	**<u>Top 10:</u> 44 records** • Hit #1 = 21 • Hit #2 = 13 • Hit #3 = 1 • Hit #4 = 5 • Hit #5 = 1 • Hit #6 = 1 • Hit #8 = 2
Hit List Record Number	**<u>Between Hit List Records #11 and #100:</u> 4 records** • #12, #18, #30, #62
Hit List Record Number	**Not in Top 100: 2 records**

Conclusion

In conclusion, the Raman prediction model has produced accurate predictions that can be used for unknown classifications and general chemical structure/functional group identification. The Wiley SmartSpectra Databases were computed based on Wiley's high quality empirical Raman spectral collection. The predicted Raman data was then vetted for outliers and records that have low validation scores. The spectrum structure validation model was used for evaluating the predicted Raman spectra and how the predictions aligned with the associated structure. This accuracy metric is valuable and therefore will be provided with the final version of the predicted library as a score for spectral matching comparison.

An external validation test was conducted to evaluate the effectiveness of the library on spectral data not included in the training set, and a hit list analysis was run on the JASCO Raman library. The records within the chemical space boundaries were computed and used as a test set. This external test resulted in a corresponding spectral first hit rate of 58% and a top ten hit rate of 82% while producing an average hit position of 13. This result was very similar but slightly lower than the model's original test set, which had a first hit rate of 64% and top ten hit rate of 91% while averaging a hit position of 6. These two test sets help illustrate the accuracy and ability of the computed library to help characterize unknown spectra.

Internal experts were also able to evaluate the library's predictions for general prediction composition and high molecular weight predictions. The latter analysis was important due to most Raman data being below 400 g/mol molecular weight. This analysis showed promise, as the higher molecular weight evaluation resulted in better hit list positioning than the JASCO test set, which confirmed that high molecular weight predicted Raman spectra would not be an issue. The internal experts also concluded that the best workflow for this library is to use it in instances where the empirical data results in low HQI, poor, or no matches. The internal experts determined that the library is also satisfactory for searching, though not as high quality as empirical data but sufficient to classify unknown spectra.

In summary, the performance level of the model's predictions validates the utility of the predicted data sets within Wiley's SmartSpectra Raman Database Collection[17] 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 Raman 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 Raman chemical space.

References

1. KnowItAll Raman Spectral Database Collection. John Wiley and Sons, Inc. 2024 https://sciencesolutions.wiley.com/solutions/technique/raman/knowitall-raman-collection/ (accessed 2024-07-18)

2. JASCO Raman Spectral Database. JASCO, Inc. 2024 JASCO Raman Library https://jascoinc.com/knowledge-base/ (accessed 2024-07-18)

3. KnowItAll Analytical Edition 2024, Release 24.0.59.0.; John Wiley and Sons, Inc. 2024. https://sciencesolutions.wiley.com/knowitall-analytical-edition-software/ (accessed 2024-07-18)

4. RDKit: Open-source cheminformatics. 2024 https://www.rdkit.org (accessed 2024-07-18)

5. Brownlee, J. (2020). A Gentle Introduction to the Rectified Linear Unit(ReLU). Machinelearningmastery.com https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/ (accessed 2024-07-18)

6. Brownlee, J. (2022). Dropout Regularization in Deep Learning Models with Keras https://machinelearningmastery.com/dropout-regularization-deep-learning-models-keras/ (accessed 2024-07-18)

7. Chollet, F., & Others (2015). Keras: Deep Learning for humans. Dense layer (keras.io) (accessed 2024-07-18)

8. Brownlee, J. (2019). Overfitting and Underfitting with Machine Learning https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/ (accessed 2024-07-18)

9. Brownlee, J. (2021). A Gentle Introduction to Sigmoid Function https://machinelearningmastery.com/a-gentle-introduction-to-sigmoid-function/ (accessed 2024-07-18)

10. Chollet, F., & Others (2015). Keras: Deep Learning for humans. https://keras.io (accessed 2024-07-18)

11. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jozefowicz, R., Jia, Y., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Schuster, M., Monga, R., Moore, S., Murray, D., Olah, C., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng., X. (2015) TensorFlow: Large-scale machine learning on heterogeneous systems. Tensorflow, tensorflow-whitepaper2015.pdf (accessed 2024-07-18)

12. 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 (accessed 2024-07-18)

13. Patino CM, Ferreira JC. Internal and external validity: can you apply research study results to your patients? *J Bras Pneumol.* **2018** *44*(3):183. doi: 10.1590/S1806-37562018000000164. PMID: 30043882; PMCID: PMC6188693.

14. John Dudovskiy. Business Research Methodology, https://research-methodology.net/research-methods/quantitative-research/correlation-regression/ (accessed 2024-07-18)

15. Brownlee, J. (2021) Random Oversampling and Undersampling for Imbalanced Classificaiton. https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/ (accessed 2024-07-18)

16. Tantra, R., McCabe, A., Bailey, M., Knight, A. and Smith, E. (2024) Comparable Raman Spectroscopy Across Different Instruments and Excitation Wavelengths. In *NPL Report DQL-AS 012,* National Physical Laboratory, Teddington, UK. https://eprintspublications.npl.co.uk/3106/1/DQL_AS12.pdf (accessed 2024-07-18)

17. KnowItAll SmartSpectra Raman Database Collection, John Wiley and Sons, Inc. 2024. https://sciencesolutions.wiley.com/solutions/technique/raman/wiley-smartspectra-raman-database-collection/ (accessed 2024-07-18)